Lecture Notes of Matrix Computations

Wen-Wei Lin

Department of Mathematics
National Tsing Hua University
Hsinchu, Taiwan 30043, R.O.C.

May 5, 2008

Contents

List of Tables

List of Figures

Part I

On the Numerical Solutions of Linear Systems

Chapter 1

Introduction

1.1 Mathematical auxiliary, definitions and relations

1.1.1 Vectors and matrices

$$A \in \mathbb{K}^{m\times n}, \text{ where } \mathbb{K} = \mathbb{R} \text{ or } \mathbb{C} \quad \Leftrightarrow \quad A = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}.$$

- Product of matrices ($\mathbb{K}^{m\times n} \times \mathbb{K}^{n\times p} \to \mathbb{K}^{m\times p}$): $C = AB$, where $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$, $i = 1, \cdots, m$, $j = 1, \cdots, p$.
- Transpose ($\mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m}$): $C = A^T$, where $c_{ij} = a_{ji} \in \mathbb{R}$.
- Conjugate transpose ($\mathbb{C}^{m\times n} \to \mathbb{C}^{n\times m}$): $C = A^*$ or $C = A^H$, where $c_{ij} = \bar{a}_{ji} \in \mathbf{C}$.
- Differentiation ($\mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$): Let $C(t) = (c_{ij}(t))$. Then $\dot{C}(t) = [\dot{c}_{ij}(t)]$.
- If $A, B \in \mathbb{K}^{n\times n}$ satisfy $AB = I$, then B is the inverse of A and is denoted by A^{-1}. If A^{-1} exists, then A is said to be nonsingular; otherwise, A is singular. A is nonsingular if and only if $\det(A) \neq 0$.
- If $A \in \mathbb{K}^{m\times n}$, $x \in \mathbb{K}^n$ and $y = Ax$, then $y_i = \sum_{j=1}^{n} a_{ij}x_j$, $i = 1, \cdots, m$.
- Outer product of $x \in \mathbb{K}^m$ and $y \in \mathbb{K}^n$:

$$xy^* = \begin{bmatrix} x_1\bar{y}_1 & \cdots & x_1\bar{y}_n \\ \vdots & \ddots & \vdots \\ x_m\bar{y}_1 & \cdots & x_m\bar{y}_n \end{bmatrix} \in \mathbb{K}^{m\times n}.$$

- Inner product of x and $y \in \mathbb{K}^n$:

$$(x, y) := x^T y = \sum_{i=1}^{n} x_i y_i = y^T x \in \mathbb{R}$$

$$(x, y) := x^* y = \sum_{i=1}^{n} \bar{x}_i y_i = \overline{y^* x} \in \mathbb{C}$$

- Sherman-Morrison Formula:
  Let $A \in \mathbb{R}^{n\times n}$ be nonsingular, u, $v \in \mathbb{R}^n$. If $v^T A^{-1}u \neq -1$, then
  $$(A + uv^T)^{-1} = A^{-1} - (1 + v^T A^{-1}u)^{-1}A^{-1}uv^T A^{-1}. \tag{1.1.1}$$

- Sherman-Morrison-Woodburg Formula:
  Let $A \in \mathbb{R}^{n\times n}$, be nonsingular U, $V \in \mathbb{R}^{n\times k}$. If $(I + V^T A^{-1}U)$ is invertible, then
  $$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1},$$

Proof of (1.1.1):

$$\begin{aligned}
&(A + uv^T)[A^{-1} - A^{-1}uv^T A^{-1}/(1 + v^T A^{-1}u)] \\
&= I + \frac{1}{1 + v^T A^{-1}u}[uv^T A^{-1}(1 + v^T A^{-1}u) - uv^T A^{-1} - uv^T A^{-1}uv^T A^{-1}] \\
&= I + \frac{1}{1 + v^T A^{-1}u}[u(v^T A^{-1}u)v^T A^{-1} - uv^T A^{-1}uv^T A^{-1}] = I.
\end{aligned}$$

■

Example 1.1.1

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

1.1.2 Rank and orthogonality

Let $A \in \mathbb{R}^{m\times n}$. Then

- $\mathcal{R}(A) = \{y \in \mathbb{R}^m \mid y = Ax \text{ for some } x \in \mathbb{R}^n \} \subseteq \mathbb{R}^m$ is the range space of A.
- $\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0 \} \subseteq \mathbb{R}^n$ is the null space of A.
- $\text{rank}(A) = \dim [\mathcal{R}(A)]$ = The number of maximal linearly independent columns of A.
- $\text{rank}(A) = \text{rank}(A^T)$.
- $\dim(\mathcal{N}(A)) + \text{rank}(A) = n$.
- If $m = n$, then A is nonsingular $\Leftrightarrow \mathcal{N}(A) = \{0\} \Leftrightarrow \text{rank}(A) = n$.
- Let $\{x_1, \cdots, x_p\} \subseteq \mathbb{R}^n$. Then $\{x_1, \cdots, x_p\}$ is said to be orthogonal if $x_i^T x_j = 0$, for $i \neq j$ and orthonormal if $x_i^T x_j = \delta_{ij}$, where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$.
- $S^\perp = \{y \in \mathbb{R}^m \mid y^T x = 0, \text{ for } x \in S\}$ = orthogonal complement of S.
- $\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A)$, $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$.
- $\mathcal{R}(A^T) \perp \mathcal{N}(A)$, $\mathcal{R}(A)^\perp = \mathcal{N}(A^T)$.

$A \in \mathbb{R}^{n\times n}$	$A \in \mathbb{C}^{n\times n}$
Symmetric: $A^T = A$	Hermitian: $A^* = A (A^H = A)$
skew-symmetric: $A^T = -A$	skew-Hermitian: $A^* = -A$
positive definite: $x^T Ax > 0, x \neq 0$	positive definite: $x^* Ax > 0, x \neq 0$
non-negative definite: $x^T Ax \geq 0$	non-negative definite: $x^* Ax \geq 0$
indefinite: $(x^T Ax)(y^T Ay) < 0$ for some x, y	indefinite: $(x^* Ax)(y^* Ay) < 0$ for some x, y
orthogonal: $A^T A = I_n$	unitary: $A^* A = I_n$
normal: $A^T A = AA^T$	normal: $A^* A = AA^*$
positive: $a_{ij} > 0$	
non-negative: $a_{ij} \geq 0$.	

Table 1.1: Some definitions for matrices.

1.1.3 Special matrices

Let $A \in \mathbb{K}^{n\times n}$. Then the matrix A is

- *diagonal* if $a_{ij} = 0$, for $i \neq j$. Denote $D = diag(d_1, \cdots, d_n) \in \mathbf{D}_n$ the set of diagonal matrices;
- *tridiagonal* if $a_{ij} = 0, |i - j| > 1$;
- *upper bi-diagonal* if $a_{ij} = 0, i > j$ or $j > i + 1$;
- *(strictly) upper triangular* if $a_{ij} = 0, i > j \quad (i \geq j)$;
- *upper Hessenberg* if $a_{ij} = 0, i > j + 1$.
  (Note: the lower case is the same as above.)

Sparse matrix: n^{1+r}, where $r < 1$ (usually between $0.2 \sim 0.5$). If $n = 1000$, $r = 0.9$, then $n^{1+r} = 501187$.

Example 1.1.2 *If S is skew-symmetric, then $I - S$ is nonsingular and $(I - S)^{-1}(I + S)$ is orthogonal (Cayley transformation of S).*

1.1.4 Eigenvalues and Eigenvectors

Definition 1.1.1 *Let $A \in \mathbb{C}^{n\times n}$. Then $\lambda \in \mathbb{C}$ is called an eigenvalue of A, if there exists $x \neq 0$, $x \in \mathbb{C}^n$ with $Ax = \lambda x$ and x is called an eigenvector corresponding to λ.*

Notations:

$\sigma(A)$:= Spectrum of A = The set of eigenvalues of A.

$\rho(A)$:= Radius of $A = \max\{|\lambda| : \lambda \in \sigma(A)\}$.

- $\lambda \in \sigma(A) \quad \Leftrightarrow \quad det(A - \lambda I) = 0$.
- $p(\lambda) = \det(\lambda I - A)$ = The characteristic polynomial of A.
- $p(\lambda) = \prod_{i=1}^{s} (\lambda - \lambda_i)^{m(\lambda_i)}$, where $\lambda_i \neq \lambda_j$ (for $i \neq j$) and $\sum_{i=1}^{s} m(\lambda_i) = n$.

- $m(\lambda_i)$ = The algebraic multiplicity of λ_i.
- $n(\lambda_i) = n - \text{rank}(A - \lambda_i I)$ = The geometric multiplicity of λ_i.
- $1 \leq n(\lambda_i) \leq m(\lambda_i)$.

Definition 1.1.2 *If there is some i such that $n(\lambda_i) < m(\lambda_i)$, then A is called degenerated.*

The following statements are equivalent:

(1) There are n linearly independent eigenvectors;

(2) A is diagonalizable, i.e., there is a nonsingular matrix T such that $T^{-1}AT \in \mathbf{D}_n$;

(3) For each $\lambda \in \sigma(A)$, it holds that $m(\lambda) = n(\lambda)$.

If A is degenerated, then eigenvectors and principal vectors derive the Jordan form of A. (See Gantmacher: Matrix Theory I, II)

Theorem 1.1.1 (Schur) (1) *Let $A \in \mathbb{C}^{n\times n}$. There is a unitary matrix U such that $U^*AU(= U^{-1}AU)$ is upper triangular.*

(2) *Let $A \in \mathbb{R}^{n\times n}$. There is an orthogonal matrix Q such that $Q^TAQ(= Q^{-1}AQ)$ is quasi-upper triangular, i.e., an upper triangular matrix possibly with nonzero subdiagonal elements in non-consecutive positions.*

(3) *A is normal if and only if there is a unitary U such that $U^*AU = D$ diagonal.*

(4) *A is Hermitian if and only if A is normal and $\sigma(A) \subseteq \mathbb{R}$.*

(5) *A is symmetric if and only if there is an orthogonal U such that $U^TAU = D$ diagonal and $\sigma(A) \subseteq \mathbb{R}$.*

1.2 Norms and eigenvalues

Let X be a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$.

Definition 1.2.1 (Vector norms) *Let N be a real-valued function defined on X ($N : X \to \mathbb{R}_+$). Then N is a (vector) norm, if*

N1: $N(\alpha x) = |\alpha| N(x)$, $\alpha \in \mathbb{K}$, *for* $x \in X$;

N2: $N(x + y) \leq N(x) + N(y)$, *for* $x, y \in X$;

N3: $N(x) = 0$ *if and only if* $x = 0$.

The usual notation is $\|x\| = N(x)$.

Example 1.2.1 *Let $X = \mathbb{C}^n$, $p \geq 1$. Then $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ is a p-norm. Especially,*

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad (\ 1\text{-}norm),$$
$$\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2} \quad (2\text{-}norm = Euclidean\text{-}norm),$$
$$\|x\|_\infty = \max_{1\leq i\leq n} |x_i| \quad (\infty\text{-}norm = maximum\ norm).$$

Lemma 1.2.1 *$N(x)$ is a continuous function in the components $x_1, \cdots, x_n$ of x.*

Proof:

$$\begin{aligned} |N(x) - N(y)| &\leq N(x-y) \leq \sum_{j=1}^n |x_j - y_j| N(e_j) \\ &\leq \|x-y\|_\infty \sum_{j=1}^n N(e_j). \end{aligned}$$

■

Theorem 1.2.1 (Equivalence of norms) *Let N and M be two norms on $\mathbb{C}^n$. Then there are constants $c_1, c_2 > 0$ such that*

$$c_1 M(x) \leq N(x) \leq c_2 M(x), \ \textit{for all } x \in \mathbf{C}^n.$$

Proof: Without loss of generality (W.L.O.G.) we can assume that $M(x) = \|x\|_\infty$ and N is arbitrary. We claim that

$$c_1 \|x\|_\infty \leq N(x) \leq c_2 \|x\|_\infty,$$

equivalently,

$$c_1 \leq N(z) \leq c_2, \forall\ z \in S = \{z \in \mathbb{C}^n | \|z\|_\infty = 1\}.$$

From Lemma 1.2.1, N is continuous on S (closed and bounded). By maximum and minimum principle, there are $c_1, c_2 \geq 0$ and $z_1, z_2 \in S$ such that

$$c_1 = N(z_1) \leq N(z) \leq N(z_2) = c_2.$$

If $c_1 = 0$, then $N(z_1) = 0$, and thus, $z_1 = 0$. This contradicts that $z_1 \in S$. ■

Remark 1.2.1 *Theorem 1.2.1 does not hold in infinite dimensional space.*

Definition 1.2.2 (Matrix-norms) *Let $A \in \mathbb{C}^{m\times n}$. A real-valued function $\|\cdot\| : \mathbf{C}^{m\times n} \to \mathbb{R}_+$ satisfying*

N1: $\|\alpha A\| = |\alpha| \|A\|$;

N2: $\|A + B\| \leq \|A\| + \|B\|$;

N3: $\|A\| = 0$ *if and only if* $A = 0$*;*

N4: $\|AB\| \le \|A\|\|B\|$ *;*

N5: $\|Ax\|_v \le \|A\|\|x\|_v$ *(matrix and vector norms are compatible for some* $\|\cdot\|_v$*)*

is called a matrix norm. If $\|\cdot\|$ *satisfies N1 to N4, then it is called a multiplicative or algebra norm.*

Example 1.2.2 (Frobenius norm) *Let* $\|A\|_F = [\sum_{i,j=1}^n |a_{i,j}|^2]^{1/2}$.

$$\begin{aligned}
\|AB\|_F &= (\sum_{i,j} |\sum_k a_{ik}b_{kj}|^2)^{\frac{1}{2}} \\
&\le (\sum_{i,j}\{\sum_k |a_{ik}|^2\}\{\sum_k |b_{kj}|^2\})^{\frac{1}{2}} \quad \textit{(Cauchy-Schwartz Ineq.)} \\
&= (\sum_i \sum_k |a_{ik}|^2)^{\frac{1}{2}} (\sum_j \sum_k |b_{kj}|^2)^{\frac{1}{2}} \\
&= \|A\|_F \|B\|_F. \qquad (1.2.1)
\end{aligned}$$

This implies that N4 holds. Furthermore, by Cauchy-Schwartz inequality we have

$$\begin{aligned}
\|Ax\|_2 &= (\sum_i |\sum_j a_{ij}x_j|^2)^{\frac{1}{2}} \\
&\le \left(\sum_i (\sum_j |a_{ij}|^2)(\sum_j |x_j|^2)\right)^{\frac{1}{2}} \\
&= \|A\|_F \|x\|_2. \qquad (1.2.2)
\end{aligned}$$

This implies that N5 holds. Also, N1, N2 and N3 hold obviously. (Here, $\|I\|_F = \sqrt{n}$*).* ■

Example 1.2.3 (Operator norm) *Given a vector norm* $\|\cdot\|$*. An associated (induced) matrix norm is defined by*

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}. \qquad (1.2.3)$$

Then N5 holds immediately. On the other hand,

$$\begin{aligned}
\|(AB)x\| &= \|A(Bx)\| \le \|A\|\|Bx\| \\
&\le \|A\|\|B\|\|x\| \qquad (1.2.4)
\end{aligned}$$

for all $x \neq 0$*. This implies that*

$$\|AB\| \le \|A\|\|B\|. \qquad (1.2.5)$$

It holds N4. (Here $\|I\| = 1$*).* ■

In the following, we represent and verify three useful matrix norms:

$$\|A\|_1 = \sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \tag{1.2.6}$$

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \tag{1.2.7}$$

$$\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(A^*A)} \tag{1.2.8}$$

Proof of (1.2.6):

$$\begin{aligned} \|Ax\|_1 &= \sum_i |\sum_j a_{ij}x_j| \leq \sum_i \sum_j |a_{ij}||x_j| \\ &= \sum_j |x_j| \sum_i |a_{ij}|. \end{aligned}$$

Let $C_1 := \sum_i |a_{ik}| = \max_j \sum_i |a_{ij}|$. Then $\|Ax\|_1 \leq C_1\|x\|_1$, thus $\|A\|_1 \leq C_1$. On the other hand, $\|e_k\|_1 = 1$ and $\|Ae_k\|_1 = \sum_{i=1}^n |a_{ik}| = C_1$. ■

Proof of (1.2.7):

$$\begin{aligned} \|Ax\|_\infty &= \max_i |\sum_j a_{ij}x_j| \\ &\leq \max_i \sum_j |a_{ij}x_j| \\ &\leq \max_i \sum_j |a_{ij}|\|x\|_\infty \\ &\equiv \sum_j |a_{kj}|\|x\|_\infty \\ &\equiv C_\infty\|x\|_\infty. \end{aligned}$$

This implies that $\|A\|_\infty \leq C_\infty$. If $A = 0$, there is nothing to prove. Assume that $A \neq 0$ and the k-th row of A is nonzero. Define $z = [z_j] \in \mathbb{C}^n$ by

$$z_j = \begin{cases} \frac{\bar{a}_{kj}}{|a_{kj}|} & \text{if} \quad a_{kj} \neq 0, \\ 1 & \text{if} \quad a_{kj} = 0. \end{cases}$$

Then $\|z\|_\infty = 1$ and $a_{kj}z_j = |a_{kj}|$, for $j = 1, \ldots, n$. It follows that

$$\|A\|_\infty \geq \|Az\|_\infty = \max_i |\sum_j a_{ij}z_j| \geq |\sum_j a_{kj}z_j| = \sum_{j=1}^{n} |a_{kj}| \equiv C_\infty.$$

Thus, $\|A\|_\infty \geq \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \equiv C_\infty$. ■

Proof of (1.2.8): Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of A^*A. There are mutually orthonormal vectors v_j, $j = 1, \ldots, n$ such that $(A^*A)v_j = \lambda_j v_j$. Let $x = \sum_j \alpha_j v_j$. Since $\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax)$,

$$\|Ax\|_2^2 = \left(\sum_j \alpha_j v_j, \sum_j \alpha_j \lambda_j v_j\right) = \sum_j \lambda_j |\alpha_j|^2 \leq \lambda_1 \|x\|_2^2.$$

Therefore, $\|A\|_2^2 \le \lambda_1$. Equality follows by choosing $x = v_1$ and $\|Av_1\|_2^2 = (v_1, \lambda_1 v_1) = \lambda_1$. So, we have $\|A\|_2 = \sqrt{\rho(A^*A)}$. ■

Example 1.2.4 (Dual norm) *Let $\frac{1}{p} + \frac{1}{q} = 1$. Then $\|\cdot\|_p^* = \|\cdot\|_q$, $(p = \infty, q = 1)$. (It concludes from the application of the Hölder inequality that $|y^*x| \le \|x\|_p \|y\|_q$.)*

Theorem 1.2.2 *Let $A \in \mathbb{C}^{n\times n}$. Then for any operator norm $\|\cdot\|$, it holds*

$$\rho(A) \le \|A\|.$$

Moreover, for any $\varepsilon > 0$, there exists an operator norm $\|\cdot\|_\varepsilon$ such that

$$\|\cdot\|_\varepsilon \le \rho(A) + \varepsilon.$$

Proof: Let $|\lambda| = \rho(A) \equiv \rho$ and x be the associated eigenvector with $\|x\| = 1$. Then,

$$\rho(A) = |\lambda| = \|\lambda x\| = \|Ax\| \le \|A\|\|x\| = \|A\|.$$

On the other hand, there is a unitary matrix U such that $A = U^*RU$, where R is upper triangular. Let $D_t = \mathrm{diag}(t, t^2, \ldots, t^n)$. Compute

$$D_t R D_t^{-1} = \begin{bmatrix} \lambda_1 & t^{-1}r_{12} & t^{-2}r_{13} & \cdots & t^{-n+1}r_{1n} \\ & \lambda_2 & t^{-1}r_{23} & \cdots & t^{-n+2}r_{2n} \\ & & \lambda_3 & & \vdots \\ & & & \ddots & t^{-1}r_{n-1,n} \\ & & & & \lambda_n \end{bmatrix}.$$

For $t > 0$ sufficiently large, the sum of all absolute values of the off-diagonal elements of $D_t R D_t^{-1}$ is less than ε. So, it holds $\|D_t R D_t^{-1}\|_1 \le \rho(A) + \varepsilon$ for sufficiently large $t(\varepsilon) > 0$. Define $\|\cdot\|_\varepsilon$ for any B by

$$\begin{aligned} \|B\|_\varepsilon &= \|D_t U B U^* D_t^{-1}\|_1 \\ &= \|(UD_t^{-1})^{-1} B (UD_t^{-1})\|_1. \end{aligned}$$

This implies that

$$\|A\|_\varepsilon = \|D_t R D_t^{-1}\| \le \rho(A) + \varepsilon.$$

■

Remark 1.2.2

$$\|UAV\|_F = \|A\|_F \quad (by\ \|UA\|_F = \sqrt{\|Ua_1\|_2^2 + \cdots + \|Ua_n\|_2^2}), \tag{1.2.9}$$

$$\|UAV\|_2 = \|A\|_2 \quad (by\ \rho(A^*A) = \rho(AA^*)), \tag{1.2.10}$$

where U and V are unitary.

Theorem 1.2.3 (Singular Value Decomposition (SVD)) *Let $A \in \mathbb{C}^{m\times n}$. Then there exist unitary matrices $U = [u_1, \cdots, u_m] \in \mathbb{C}^{m\times m}$ and $V = [v_1, \cdots, v_n] \in \mathbb{C}^{n\times n}$ such that*

$$U^*AV = diag(\sigma_1, \cdots, \sigma_p) = \Sigma,$$

where $p = \min\{m, n\}$ and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p \ge 0$. (Here, σ_i denotes the i-th largest singular value of A).

Proof: There are $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$ with $\|x\|_2 = \|y\|_2 = 1$ such that $Ax = \sigma y$, where $\sigma = \|A\|_2$ ($\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$). Let $V = [x, V_1] \in \mathbb{C}^{n\times n}$, and $U = [y, U_1] \in \mathbb{C}^{m\times m}$ be unitary. Then

$$A_1 \equiv U^*AV = \begin{bmatrix} \sigma & w^* \\ 0 & B \end{bmatrix}.$$

Since

$$\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 \geq (\sigma^2 + w^*w)^2,$$

it follows that

$$\|A_1\|_2^2 \geq \sigma^2 + w^*w \quad \text{from} \quad \frac{\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2}{\left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2} \geq \sigma^2 + w^*w.$$

But $\sigma^2 = \|A\|_2^2 = \|A_1\|_2^2$, it implies $w = 0$. Hence, the theorem holds by induction. ■

Remark 1.2.3 $\|A\|_2 = \sqrt{\rho(A^*A)} = \sigma_1 =$ *The maximal singular value of* A.

Let $A = U\Sigma V^*$. Then we have

$$\begin{aligned} \|ABC\|_F &= \|U\Sigma V^*BC\|_F = \|\Sigma V^*BC\|_F \\ &\leq \sigma_1 \|BC\|_F = \|A\|_2 \|BC\|_F. \end{aligned}$$

This implies

$$\|ABC\|_F \leq \|A\|_2 \|B\|_F \|C\|_2. \tag{1.2.11}$$

In addition, by (1.2.2) and (1.2.11), we get

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2. \tag{1.2.12}$$

Theorem 1.2.4 *Let* $A \in \mathbb{C}^{n\times n}$. *The following statements are equivalent:*
(1) $\lim_{m\to\infty} A^m = 0$;
(2) $\lim_{m\to\infty} A^m x = 0$ *for all* x;
(3) $\rho(A) < 1$.

Proof: (1) $\Rightarrow$ (2): Trivial. (2) $\Rightarrow$ (3): Let $\lambda \in \sigma(A)$, i.e., $Ax = \lambda x$, $x \neq 0$. This implies $A^m x = \lambda^m x \to 0$, as $\lambda^m \to 0$. Thus $|\lambda| < 1$, i.e., $\rho(A) < 1$. (3) $\Rightarrow$ (1): There is a norm $\|\cdot\|$ with $\|A\| < 1$ (by Theorem 1.2.2). Therefore, $\|A^m\| \leq \|A\|^m \to 0$, i.e., $A^m \to 0$. ■

Theorem 1.2.5 *It holds that*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$$

where $\|\ \|$ *is an operator norm.*

Proof: Since

$$\rho(A)^k = \rho(A^k) \leq \|A^k\| \;\Rightarrow\; \rho(A) \leq \|A^k\|^{1/k},$$

for $k = 1, 2, \ldots$. If $\varepsilon > 0$, then $\tilde{A} = [\rho(A) + \varepsilon]^{-1} A$ has spectral radius < 1 and by Theorem 1.2.4 we have $\|\tilde{A}^k\| \to 0$ as $k \to \infty$. There is an $N = N(\varepsilon, A)$ such that $\|\tilde{A}^k\| < 1$ for all $k \geq N$. Thus, $\|A^k\| \leq [\rho(A) + \varepsilon]^k$, for all $k \geq N$ or $\|A^k\|^{1/k} \leq \rho(A) + \varepsilon$ for all $k \geq N$. Since $\rho(A) \leq \|A^k\|^{1/k}$, and k, ε are arbitrary, $\lim_{k\to\infty} \|A^k\|^{1/k}$ exists and equals $\rho(A)$. ■

Theorem 1.2.6 *Let $A \in \mathbb{C}^{n\times n}$, and $\rho(A) < 1$. Then $(I - A)^{-1}$ exists and*

$$(I - A)^{-1} = I + A + A^2 + \cdots .$$

Proof: Since $\rho(A) < 1$, the eigenvalues of $(I - A)$ are nonzero. Therefore, by Theorem 2.5, $(I - A)^{-1}$ exists and

$$(I - A)(I + A + A^2 + \cdots + A^m) = I - A^m \to 0.$$

■

Corollary 1.2.1 *If $\|A\| < 1$, then $(I - A)^{-1}$ exists and*

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

Proof: Since $\rho(A) \leq \|A\| < 1$ (by Theorem 1.2.2),

$$\|(I - A)^{-1}\| = \|\sum_{i=0}^{\infty} A^i\| \leq \sum_{i=0}^{\infty} \|A\|^i = (1 - \|A\|)^{-1}.$$

■

Theorem 1.2.7 *(Without proof) For $A \in \mathbb{K}^{n\times n}$ the following statements are equivalent:*

(1) *There is a multiplicative norm p with $p(A^k) \leq 1, k = 1, 2, \ldots$.*

(2) *For each multiplicative norm p the power $p(A^k)$ are uniformly bounded, i.e., there exists a $M(p) < \infty$ such that $p(A^k) \leq M(p)$, $k = 0, 1, 2, \ldots$.*

(3) *$\rho(A) \leq 1$ and all eigenvalue λ with $|\lambda| = 1$ are not degenerated. (i.e., $m(\lambda) = n(\lambda)$.)*

(See Householder's book: *The theory of matrix,* pp.45-47.)

In the following we prove some important inequalities of vector norms and matrix norms.

(a) It holds that

$$1 \leq \frac{\|x\|_p}{\|x\|_q} \leq n^{(q-p)/pq}, \quad (p \leq q). \tag{1.2.13}$$

Proof: Claim $\|x\|_q \leq \|x\|_p$, $(p \leq q)$: It holds

$$\|x\|_q = \left\| \|x\|_p \frac{x}{\|x\|_p} \right\|_q = \|x\|_p \left\| \frac{x}{\|x\|_p} \right\|_q \leq \mathcal{C}_{p,q} \|x\|_p,$$

where

$$\mathcal{C}_{p,q} = \max_{\|e\|_p = 1} \|e\|_q, \quad e = (e_1, \cdots, e_n)^T.$$

We now show that $\mathcal{C}_{p,q} \leq 1$. From $p \leq q$, we have

$$\|e\|_q^q = \sum_{i=1}^n |e_i|^q \leq \sum_{i=1}^n |e_i|^p = 1 \quad (\text{by } |e_i| \leq 1).$$

Hence, $\mathcal{C}_{p,q} \leq 1$, thus $\|x\|_q \leq \|x\|_p$.

To prove the second inequality: Let $\alpha = q/p > 1$. Then the Jensen inequality holds for the convex function $\varphi(x)$:

$$\varphi(\int_\Omega f d\mu) \leq \int_\Omega (\varphi \circ f) d\mu, \quad \mu(\Omega) = 1.$$

If we take $\varphi(x) = x^\alpha$, then we have

$$\int_\Omega |f|^q dx = \int_\Omega (|f|^p)^{q/p} \, dx \geq \left(\int_\Omega |f|^p dx \right)^{q/p}$$

with $|\Omega| = 1$. Consider the discrete measure $\sum_{i=1}^n \frac{1}{n} = 1$ and $f(i) = |x_i|$. It follows that

$$\sum_{i=1}^n |x_i|^q \frac{1}{n} \geq \left(\sum_{i=1}^n |x_i|^p \frac{1}{n} \right)^{q/p}.$$

Hence, we have

$$n^{-\frac{1}{q}} \|x\|_q \geq n^{-\frac{1}{p}} \|x\|_p.$$

Thus,

$$n^{(q-p)/pq} \|x\|_q \geq \|x\|_p.$$

(b) It holds that

$$1 \leq \frac{\|x\|_p}{\|x\|_\infty} \leq n^{\frac{1}{p}}. \tag{1.2.14}$$

Proof: Let $q \to \infty$ and $\lim_{q \to \infty} \|x\|_q = \|x\|_\infty$:

$$\|x\|_\infty = |x_k| = (|x_k|^q)^{\frac{1}{q}} \leq \left(\sum_{i=1}^n |x_i|^q \right)^{\frac{1}{q}} = \|x\|_q.$$

On the other hand, we have

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q\right)^{\frac{1}{q}} \leq (n\|x\|_\infty^q)^{\frac{1}{q}} \leq n^{\frac{1}{q}}\|x\|_\infty$$

which implies that $\lim_{q\to\infty} \|x\|_q = \|x\|_\infty$.

(c) It holds that

$$\max_{1\leq j\leq n} \|a_j\|_p \leq \|A\|_p \leq n^{(p-1)/p} \max_{1\leq j\leq n} \|a_j\|_p, \tag{1.2.15}$$

where $A = [a_1, \cdots, a_n] \in \mathbb{R}^{m\times n}$.

Proof: The first inequality holds obviously. To show the second inequality, for $\|y\|_p = 1$ we have

$$\begin{aligned} \|Ay\|_p &\leq \sum_{j=1}^{n} |y_j|\|a_j\|_p \leq \sum_{j=1}^{n} |y_j| \max_j \|a_j\|_p \\ &= \|y\|_1 \max_j \|a_j\|_p \leq n^{(p-1)/p} \max_j \|a_j\|_p \quad \text{(by (1.2.13))}. \end{aligned}$$

(d) It holds that

$$\max_{i,j} |a_{ij}| \leq \|A\|_p \leq n^{(p-1)/p} m^{1/p} \max_{i,j} |a_{ij}|, \tag{1.2.16}$$

where $A \in \mathbb{R}^{m\times n}$.

Proof: By (1.2.14) and (1.2.15) immediately.

(e) It holds that

$$m^{(1-p)/p}\|A\|_1 \leq \|A\|_p \leq n^{(p-1)/p}\|A\|_1. \tag{1.2.17}$$

Proof: By (1.2.15) and (1.2.13) immediately.

(f) By *Hölder inequality*, we have (see Appendix later!)

$$|y^* x| \leq \|x\|_p \|y\|_q,$$

where $\frac{1}{p} + \frac{1}{q} = 1$ or

$$\max\{|x^* y| \,:\, \|y\|_q = 1\} = \|x\|_p. \tag{1.2.18}$$

Then it holds that

$$\|A\|_p = \|A^T\|_q. \tag{1.2.19}$$

Proof: By (1.2.18) we have

$$\begin{aligned} \max_{\|x\|_p=1} \|Ax\|_p &= \max_{\|x\|_p=1} \max_{\|y\|_q=1} |(Ax)^T y| \\ &= \max_{\|y\|_q=1} \max_{\|x\|_p=1} |x^T(A^T y)| = \max_{\|y\|_q=1} \|A^T y\|_q = \|A^T\|_q. \end{aligned}$$

(g) It holds that

$$n^{-\frac{1}{p}}\|A\|_\infty \le \|A\|_p \le m^{\frac{1}{p}}\|A\|_\infty. \tag{1.2.20}$$

Proof: By (1.2.17) and (1.2.19), we get

$$\begin{aligned} m^{\frac{1}{p}}\|A\|_\infty &= m^{\frac{1}{p}}\|A^T\|_1 = m^{1-\frac{1}{q}}\|A^T\|_1 \\ &= m^{(q-1)/q}\|A^T\|_1 \ge \|A^T\|_q = \|A\|_p. \end{aligned}$$

(h) It holds that

$$\|A\|_2 \le \sqrt{\|A\|_p\|A\|_q}, \quad (\frac{1}{p}+\frac{1}{q}=1). \tag{1.2.21}$$

Proof: By (1.2.19) we have

$$\|A\|_p\|A\|_q = \|A^T\|_q\|A\|_q \ge \|A^TA\|_q \ge \|A^TA\|_2.$$

The last inequality holds by the following statement: Let S be a symmetric matrix. Then $\|S\|_2 \le \|S\|$, for any matrix operator norm $\|\ \|$. Since $|\lambda| \le \|S\|$,

$$\|S\|_2 = \sqrt{\rho(S^*S)} = \sqrt{\rho(S^2)} = \max_{\lambda\in\sigma(S)}|\lambda| = |\lambda_{\max}|.$$

This implies, $\|S\|_2 \le \|S\|$.

(i) For $A \in \mathbb{R}^{m\times n}$ and $q \ge p \ge 1$, it holds that

$$n^{(p-q)/pq}\|A\|_q \le \|A\|_p \le m^{(q-p)/pq}\|A\|_q. \tag{1.2.22}$$

Proof: By (1.2.13), we get

$$\begin{aligned} \|A\|_p &= \max_{\|x\|_p=1}\|Ax\|_p \le \max_{\|x\|_q\le 1} m^{(q-p)/pq}\|Ax\|_q \\ &= m^{(q-p)/pq}\|A\|_q. \end{aligned}$$

Appendix: To show Hölder inequality and (1.2.18)

Taking $\varphi(x) = e^x$ in Jensen's inequality we have

$$\exp\left\{\int_\Omega f d\mu\right\} \le \int_\Omega e^f d\mu.$$

Let Ω = finite set $= \{p_1,\dots,p_n\}$, $\mu(\{p_i\}) = \frac{1}{n}$, $f(p_i) = x_i$. Then

$$\exp\left\{\frac{1}{n}(x_1+\cdots+x_n)\right\} \le \frac{1}{n}\left(e^{x_1}+\cdots+e^{x_n}\right).$$

Taking $y_i = e^{x_i}$, we have

$$(y_1\cdots y_n)^{1/n} \le \frac{1}{n}(y_1+\cdots+y_n).$$

Taking $\mu(\{p_i\}) = q_i > 0$, $\sum_{i=1}^n q_i = 1$ we have

$$y_1^{q_1} \cdots y_n^{q_n} \le q_1 y_1 + \cdots + q_n y_n. \tag{1.2.23}$$

Let $\alpha_i = x_i/\|x\|_p$, $\beta_i = y_i/\|y\|_q$, where $x = [x_1, \cdots, x_n]^T$, $y = [y_1, \cdots, y_n]^T$, $\alpha = [\alpha_1, \cdots, \alpha_n]^T$ and $\beta = [\beta_1, \cdots, \beta_n]^T$. By (1.2.23) we have

$$\alpha_i \beta_i \le \frac{1}{p}\alpha_i^p + \frac{1}{q}\beta_i^q.$$

Since $\|\alpha\|_p = 1$, $\|\beta\|_q = 1$, it holds

$$\sum_{i=1}^n \alpha_i \beta_i \le \frac{1}{p} + \frac{1}{q} = 1.$$

Thus,

$$|x^T y| \le \|x\|_p \|y\|_q.$$

To show $\max\{|x^T y|; \|x\|_p = 1\} = \|y\|_q$. Taking $x_i = y_i^{q-1}/\|y\|_q^{q/p}$ we have

$$\|x\|_p^p = \frac{\sum_{i=1}^n |y_i|^{(q-1)p}}{\|y\|_q^q} = 1.$$

Note $(q-1)p = q$. Then

$$\left|\sum_{i=1}^n x_i^T y_i\right| = \frac{\sum_{i=1}^n |y_i|^q}{\|y\|_q^{q/p}} = \frac{\|y\|_q^q}{\|y\|_q^{q/p}} = \|y\|_q.$$

The following two properties are useful in the following sections.

(i) There exists $\hat{z}$ with $\|\hat{z}\|_p = 1$ such that $\|y\|_q = \hat{z}^T y$. Let $z = \hat{z}/\|y\|_q$. Then we have $z^T y = 1$ and $\|z\|_p = \frac{1}{\|y\|_q}$.

(ii) From the duality, we have $\|y\| = (\|y\|_*)_* = \max_{\|u\|_*=1} |y^T u| = y^T \hat{z}$ and $\|\hat{z}\|_* = 1$. Let $z = \hat{z}/\|y\|$. Then we have $z^T y = 1$ and $\|z\|_* = \frac{1}{\|y\|}$.

1.3 The Sensitivity of Linear System $Ax = b$

1.3.1 Backward error and Forward error

Let $x = F(a)$. We define backward and forward errors in Figure 1.1. In Figure 1.1, $\hat{x} + \Delta x = F(a + \Delta a)$ is called a mixed forward-backward error, where $|\Delta x| \le \varepsilon |x|$, $|\Delta a| \le \eta |a|$.

Definition 1.3.1 (i) *An algorithm is backward stable, if for all a, it produces a computed $\hat{x}$ with a small backward error, i.e., $\hat{x} = F(a + \Delta a)$ with Δa small.*

(ii) *An algorithm is numerical stable, if it is stable in the mixed forward-backward error sense, i.e., $\hat{x} + \Delta x = F(a + \Delta a)$ with both Δa and Δx small.*

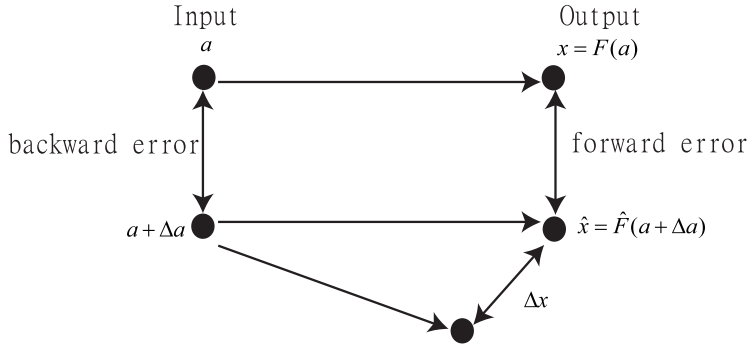


Figure 1.1: Relationship between backward and forward errors.

(iii) *If a method which produces answers with forward errors of similar magnitude to those produced by a backward stable method, is called a forward stable.*

Remark 1.3.1 **(i)** *Backward stable ⇒ forward stable, not vice versa!*

(ii) *Forward error ≤ condition number × backward error*

Consider

$$\hat{x} - x = F(a + \Delta a) - F(a) = F'(a)\Delta a + \frac{F''(a + \theta\Delta a)}{2}(\Delta a)^2,\ \theta \in (0, 1).$$

Then we have

$$\frac{\hat{x} - x}{x} = \left(\frac{aF'(a)}{F(a)}\right)\frac{\Delta a}{a} + O\left((\Delta a)^2\right).$$

The quantity $C(a) = \left|\frac{aF'(a)}{F(a)}\right|$ is called the condition number of F. If x or F is a vector, then the condition number is defined in a similar way using norms and it measures the maximum relative change, which is attained for some, but not all Δa.

$$\begin{cases} \text{À priori error estimate !} \\ \text{À posteriori error estimate !} \end{cases}$$

1.3.2 An SVD Analysis

Let $A = \sum_{i=1}^{n} \sigma_i u_i {v_i}^T = U\Sigma V^T$ be a singular value decomposition (SVD) of A. Then

$$x = A^{-1}b = (U\Sigma V^T)^{-1}b = \sum_{i=1}^{n} \frac{{u_i}^T b}{\sigma_i} v_i.$$

If $cos(\theta) =\mid {u_n}^T b \mid / \parallel b \parallel_2$ and

$$(A - \varepsilon u_n {v_n}^T)y = b + \varepsilon({u_n}^T b)u_n, \quad \sigma_n > \varepsilon \geq 0.$$

Then we have

$$\| y - x \|_2 \geq (\frac{\varepsilon}{\sigma_n}) \| x \|_2 \, cos(\theta).$$

Let $E = \text{diag}\{0, \cdots, 0, \varepsilon\}$. Then it holds

$$(\Sigma - E)V^T y = U^T b + \varepsilon({u_n}^T b)e_n.$$

Therefore,

$$\begin{aligned}
y - x &= V(\Sigma - E)^{-1}U^T b + \varepsilon({u_n}^T b)(\sigma_n - \varepsilon)^{-1}v_n - V\Sigma^{-1}U^T b \\
&= V((\Sigma - E)^{-1} - \Sigma^{-1})U^T b + \varepsilon({u_n}^T b)(\sigma_n - \varepsilon)^{-1}v_n \\
&= V(\Sigma^{-1}E(\Sigma - E)^{-1})U^T b + \varepsilon({u_n}^T b)(\sigma_n - \varepsilon)^{-1}v_n \\
&= V\text{diag}\left(0, \cdots, 0, \frac{\varepsilon}{\sigma_n(\sigma_n - \varepsilon)}\right) U^T b + \varepsilon({u_n}^T b)(\sigma_n - \varepsilon)^{-1}v_n \\
&= \frac{\varepsilon}{\sigma_n(\sigma_n - \varepsilon)} v_n({u_n}^T b) + \varepsilon({u_n}^T b)(\sigma_n - \varepsilon)^{-1}v_n \\
&= {u_n}^T b v_n(\frac{\varepsilon}{\sigma_n(\sigma_n - \varepsilon)} + \varepsilon(\sigma_n - \varepsilon)^{-1}) \\
&= \frac{\varepsilon(1 + \sigma_n)}{\sigma_n(\sigma_n - \varepsilon)} {u_n}^T b v_n.
\end{aligned}$$

From the inequality $\|x\|_2 \leq \|b\|_2 \|A^{-1}\|_2$ we have

$$\frac{\| y - x \|_2}{\| x \|_2} \geq \frac{| {u_n}^T b | \, \frac{\varepsilon}{\sigma_n}(\frac{1+\sigma}{\sigma - \varepsilon})}{\| b \|_2} \geq \frac{| {u_n}^T b |}{\| b \|_2} \frac{\varepsilon}{\sigma_n}.$$

■

Theorem 1.3.1 *A is nonsingular and $\| A^{-1}E \| = r < 1$. Then $A + E$ is nonsingular and $\| (A + E)^{-1} - A^{-1} \| \leq \| E \| \, \| A^{-1} \|^2 / (1 - r)$.*

Proof:: Since A is nonsingular, $A+E = A(I-F)$, where $F = -A^{-1}E$. Since $\| F \| = r < 1$, it follows that $I - F$ is nonsingular (by Corollary 1.2.1) and $\| (I - F)^{-1} \| < \frac{1}{1-r}$. Then

$$(A + E)^{-1} = (I - F)^{-1}A^{-1} \Longrightarrow \| (A + E)^{-1} \| \leq \frac{\|A^{-1}\|}{1 - r}$$

and

$$(A + E)^{-1} - A^{-1} = -A^{-1}E(A + E)^{-1}.$$

It follows that

$$\| (A + E)^{-1} - A^{-1} \| \leq \| A^{-1} \| \| E \| \| (A + E)^{-1} \| \leq \frac{\| A^{-1} \|^2 \| E \|}{1 - r}.$$

■

Lemma 1.3.1 *Let*

$$\begin{cases} Ax = b, \\ (A + \Delta A)y = b + \Delta b, \end{cases}$$

where $\| \Delta A \| \leq \delta \| A \|$ and $\| \Delta b \| \leq \delta \|b\|$. If $\delta\kappa(A) = r < 1$, then $A+\Delta A$ is nonsingular and $\frac{\|y\|}{\|x\|} \leq \frac{1+r}{1-r}$, where $\kappa(A) = \|A\|\|A^{-1}\|$.

Proof: Since $\| A^{-1}\Delta A \| < \delta\|A^{-1}\|\|A\| = r < 1$, it follows that $A + \Delta A$ is nonsingular. From the equality $(I + A^{-1}\Delta A)y = x + A^{-1}\Delta b$ follows that

$$\begin{aligned} \|y\| &\leq \| (I + A^{-1}\Delta A)^{-1} \| (\|x\| + \delta\|A^{-1}\|\|b\|) \\ &\leq \frac{1}{1-r}(\|x\| + \delta\|A^{-1}\|\|b\|) \\ &= \frac{1}{1-r}(\|x\| + r\frac{\| b \|}{\|A\|}). \end{aligned}$$

From $\|b\| = \| Ax \| \leq \|A\|\|x\|$ follows the lemma. ■

1.3.3 Normwise Forward Error Bound

Theorem 1.3.2 *If the assumption of Lemma 1.3.1 holds, then* $\frac{\|x-y\|}{\|x\|} \leq \frac{2\delta}{1-r}\kappa(A)$.

Proof:: Since $y - x = A^{-1}\Delta b - A^{-1}\Delta A y$, we have

$$\| y - x \| \leq \delta\|A^{-1}\|\|b\| + \delta\|A^{-1}\|\|A\|\|y\|.$$

So by Lemma 1.3.1 it holds

$$\begin{aligned} \frac{\| y - x \|}{\|x\|} &\leq \delta\kappa(A)\frac{\|b\|}{\|A\|\|x\|} + \delta\kappa(A)\frac{\|y\|}{\|x\|} \\ &\leq \delta\kappa(A)(1 + \frac{1+r}{1-r}) = \frac{2\delta}{1-r}\kappa(A). \end{aligned}$$

■

1.3.4 Componentwise Forward Error Bound

Theorem 1.3.3 *Let* $Ax = b$ *and* $(A + \Delta A)y = b + \Delta b$, *where* $| \Delta A | \leq \delta | A |$ *and* $| \Delta b | \leq \delta | b |$. *If* $\delta\kappa_\infty(A) = r < 1$, *then* $(A + \Delta A)$ *is nonsingular and* $\frac{\|y-x\|_\infty}{\|x\|_\infty} \leq \frac{2\delta}{1-r} \| | A^{-1} | | A | \|_\infty$. *Here* $\| | A^{-1} | | A | \|_\infty$ *is called a Skeel condition number.*

Proof:: Since $\| \Delta A \|_\infty \leq \delta\|A\|_\infty$ and $\| \Delta b \|_\infty \leq \delta\|b\|_\infty$, the assumptions of Lemma 1.3.1 are satisfied in ∞-norm. So, $A + \Delta A$ is nonsingular and $\frac{\|y\|_\infty}{\|x\|_\infty} \leq \frac{1+r}{1-r}$.

Since $y - x = A^{-1}\Delta b - A^{-1}\Delta A y$, we have

$$\begin{aligned} | y - x | &\leq | A^{-1} | | \Delta b | + | A^{-1} | | \Delta A | | y | \\ &\leq \delta | A^{-1} | | b | + \delta | A^{-1} | | A | | y | \\ &\leq \delta | A^{-1} | | A | (| x | + | y |). \end{aligned}$$

By taking ∞-norm, we have

$$\begin{aligned} \| y - x \|_\infty &\leq \delta \| | A^{-1} | | A | \|_\infty (\|x\|_\infty + \frac{1+r}{1-r}\|x\|_\infty) \\ &= \frac{2\delta}{1-r}\| | A^{-1} | | A | \|_\infty. \end{aligned}$$

■

1.3.5 Derivation of Condition Number of $Ax = b$

Let

$$(A + \varepsilon F)x(\varepsilon) = b + \varepsilon f \quad \text{with} \quad x(0) = x.$$

Then we have $\dot{x}(0) = A^{-1}(f - Fx)$ and $x(\varepsilon) = x + \varepsilon\dot{x}(0) + o(\varepsilon^2)$. Therefore,

$$\frac{\| x(\varepsilon) - x \|}{\|x\|} \leq \varepsilon\|A^{-1}\|\{\frac{\| f \|}{\|x\|} + \| F \|\} + o(\varepsilon^2).$$

Define condition number $\kappa(A) := \|A\|\|A^{-1}\|$. Then we have

$$\frac{\| x(\varepsilon) - x \|}{\|x\|} \leq \kappa(A)(\rho_A + \rho_b) + o(\varepsilon^2),$$

where $\rho_A = \varepsilon\|F\|/\|A\|$ and $\rho_b = \varepsilon\|f\|/\|b\|$.

1.3.6 Normwise Backward Error

Theorem 1.3.4 *Let y be the computed solution of $Ax = b$. Then the normwise backward error bound*

$$\eta(y) := \min\left\{\varepsilon | (A + \Delta A)y = b + \Delta b, \quad \|\Delta A\| \leq \varepsilon\|A\|, \quad \|\Delta b\| \leq \varepsilon\|b\|\right\}$$

is given by

$$\eta(y) = \frac{\|r\|}{\|A\|\|y\| + \|b\|}, \tag{1.3.24}$$

where $r = b - Ay$ is the residual.

Proof: The right hand side of (1.3.24) is a upper bound of $\eta(y)$. This upper bound is attained for the perturbation (by construction!)

$$\Delta A_{\min} = \frac{\|A\|\|y\|rz^T}{\|A\|\|y\| + \|b\|}, \quad \Delta b_{\min} = -\frac{\|b\|}{\|A\|\|y\| + \|b\|}r,$$

where z is the dual vector of y, i.e. $z^T y = 1$ and $\|z\|_* = \frac{1}{\|y\|}$.

Check:

$$\|\Delta A_{\min}\| = \eta(y)\|A\|,$$

or

$$\|\Delta A_{\min}\| = \frac{\|A\|\|y\|\|rz^T\|}{\|A\|\|y\| + \|b\|} = \left(\frac{\|r\|}{\|A\|\|y\| + \|b\|}\right)\|A\|.$$

That is, to prove

$$\|rz^T\| = \frac{\|r\|}{\|y\|}.$$

Since

$$\|rz^T\| = \max_{\|u\|=1}\|(rz^T)u\| = \|r\|\max_{\|u\|=1}|z^T u| = \|r\|\|z\|_* = \|r\|\frac{1}{\|y\|},$$

we have done. Similarly, $\|\Delta b_{\min}\| = \eta(y)\|b\|$. ■

1.3.7 Componentwise Backward Error

Theorem 1.3.5 *The componentwise backward error bound*

$$\omega(y) := \min\left\{\varepsilon | (A + \Delta A)y = b + \Delta b, \quad |\Delta A| \le \varepsilon |A|, \quad |\Delta b| \le \varepsilon |b|\right\}$$

is given by

$$\omega(y) = \max_i \frac{|r|_i}{(A|y| + b)_i}, \tag{1.3.25}$$

where $r = b - Ay$. (note: $\xi/0 = 0$ if $\xi = 0$; $\xi/0 = \infty$ if $\xi \neq 0$.)

Proof: The right hand side of (1.3.25) is a upper bound for $\omega(y)$. This bound is attained for the perturbations $\Delta A = D_1 A D_2$ and $\Delta b = -D_1 b$, where $D_1 = \mathrm{diag}(r_i/(A|y| + b)_i)$ and $D_2 = \mathrm{diag}(\mathrm{sign}(y_i))$. ■

Remark 1.3.2 *Theorems 1.3.4 and 1.3.5 are posterior error estimation approach.*

1.3.8 Determinants and Nearness to Singularity

$$B_n = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ & 1 & \ddots & \vdots \\ & & 1 & -1 \\ 0 & & & 1 \end{bmatrix}, \quad B_n^{-1} = \begin{bmatrix} 1 & 1 & \cdots & 2^{n-2} \\ & \ddots & \ddots & \vdots \\ & & \ddots & 1 \\ 0 & & & 1 \end{bmatrix}.$$

Then $\det(B_n) = 1$, $\kappa_\infty(B_n) = n2^{n-1}$, $\sigma_{30}(B_{30}) \approx 10^{-8}$.

$$D_n = \begin{bmatrix} 10^{-1} & & 0 \\ & \ddots & \\ 0 & & 10^{-1} \end{bmatrix}.$$

Then $\det(D_n) = 10^{-n}$, $\kappa_p(D_n) = 1$ and $\sigma_n(D_n) = 10^{-1}$.

Chapter 2

Numerical methods for solving linear systems

Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix. We want to solve the linear system $Ax = b$ by (a) Direct methods (finite steps); Iterative methods (convergence). (See Chapter 4)

2.1 Elementary matrices

Let $\mathbb{X} = \mathbb{K}^n$ and $x, y \in \mathbb{X}$. Then $y^*x \in \mathbb{K}$, $xy^* = \begin{pmatrix} x_1\bar{y_1} & \cdots & x_1\bar{y_n} \\ \vdots & & \vdots \\ x_n\bar{y_1} & \cdots & x_n\bar{y_n} \end{pmatrix}$. The eigenvalues of xy^* are $\{0, \cdots, 0, y^*x\}$, since $\text{rank}(xy^*) = 1$ by $(xy^*)z = (y^*z)x$ and $(xy^*)x = (y^*x)x$.

Definition 2.1.1 *A matrix of the form*

$$I - \alpha xy^* \quad (\alpha \in \mathbb{K}, x, y \in \mathbb{K}^n) \tag{2.1.1}$$

is called an elementary matrix.

The eigenvalues of $(I - \alpha xy^*)$ are $\{1, 1, \cdots, 1, 1 - \alpha y^*x\}$. Compute

$$(I - \alpha xy^*)(I - \beta xy^*) = I - (\alpha + \beta - \alpha\beta y^*x)xy^*. \tag{2.1.2}$$

If $\alpha y^*x - 1 \neq 0$ and let$\beta = \frac{\alpha}{\alpha y^*x - 1}$, then $\alpha + \beta - \alpha\beta y^*x = 0$. We have

$$(I - \alpha xy^*)^{-1} = (I - \beta xy^*),\ \frac{1}{\alpha} + \frac{1}{\beta} = y^*x. \tag{2.1.3}$$

Example 2.1.1 *Let $x \in \mathbb{K}^n$, and $x^*x = 1$. Let $H = \{z :\ z^*x = 0\}$ and*

$$Q = I - 2xx^* \quad (Q = Q^*, Q^{-1} = Q).$$

Then Q reflects each vector with respect to the hyperplane H. Let $y = \alpha x + w$, $w \in H$. Then, we have

$$Qy = \alpha Qx + Qw = -\alpha x + w - 2(x^*w)x = -\alpha x + w.$$

Example 2.1.2 *Let $y = e_i$ = the i-th column of unit matrix and $x = l_i = [0, \cdots, 0, l_{i+1,i}, \cdots, l_{n,i}]^T$. Then,*

$$I + l_i e_i^T = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & l_{i+1,i} & \\ & & \vdots & \ddots \\ & & l_{n,i} & & 1 \end{bmatrix} \tag{2.1.4}$$

Since $e_i^T l_i = 0$, we have

$$(I + l_i e_i^T)^{-1} = (I - l_i e_i^T). \tag{2.1.5}$$

From the equality

$$(I + l_1 e_1^T)(I + l_2 e_2^T) = I + l_1 e_1^T + l_2 e_2^T + l_1 (e_1^T l_2) e_2^T = I + l_1 e_1^T + l_2 e_2^T$$

follows that

$$\begin{aligned} (I + l_1 e_1^T) \cdots (I + l_i e_i^T) \cdots (I + l_{n-1} e_{n-1}^T) &= I + l_1 e_1^T + l_2 e_2^T + \cdots + l_{n-1} e_{n-1}^T \\ &= \begin{bmatrix} 1 & & & \\ l_{21} & \ddots & 0 & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix}. \end{aligned} \tag{2.1.6}$$

Theorem 2.1.1 *A lower triangular with "1" on the diagonal can be written as the product of $n-1$ elementary matrices of the form (2.1.4).*

Remark 2.1.1 $(I + l_1 e_1^T + \ldots + l_{n-1} e_{n-1}^T)^{-1} = (l - l_{n-1} e_{n-1}^T) \ldots (I - l_1 e_1^T)$ *which can not be simplified as in (2.1.6).*

2.2 LR-factorization

Definition 2.2.1 *Given $A \in \mathbb{C}^{n \times n}$, a lower triangular matrix L and an upper triangular matrix R. If $A = LR$, then the product LR is called a LR-factorization (or LR-decomposition) of A.*

Basic problem:

Given $b \neq 0$, $b \in \mathbb{K}^n$. Find a vector $l_1 = [0, l_{21}, \ldots, l_{n1}]^T$ and $c \in \mathbb{K}$ such that

$$(I - l_1 e_1^T) b = c e_1.$$

Solution:

$$\begin{cases} b_1 = c, \\ b_i - l_{i1} b_1 = 0, \quad i = 2, \ldots, n. \end{cases}$$

$$\begin{cases} b_1 = 0, & \text{it has no solution (since } b \neq 0), \\ b_1 \neq 0, & \text{then } c = b_1,\ l_{i1} = b_i / b_1,\ i = 2, \ldots, n. \end{cases}$$

Construction of LR-factorization:

Let $A = A^{(0)} = [a_1^{(0)} \mid \ldots \mid a_n^{(0)}]$. Apply basic problem to $a_1^{(0)}$: If $a_{11}^{(0)} \neq 0$, then there exists $L_1 = I - l_1 e_1^T$ such that $(I - l_1 e_1^T) a_1^{(0)} = a_{11}^{(0)} e_1$. Thus

$$A^{(1)} = L_1 A^{(0)} = [L_1 a_1^{(0)} \mid \ldots \mid L_1 a_n^{(0)}] = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \ldots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} \end{bmatrix}. \tag{2.2.1}$$

The i-th step:

$$\begin{aligned} A^{(i)} &= L_i A^{(i-1)} = L_i L_{i-1} \ldots L_1 A^{(0)} \\ &= \begin{bmatrix} a_{11}^{(0)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \cdots & \cdots & \cdots & a_{2n}^{(1)} \\ \vdots & 0 & \ddots & & & \vdots \\ \vdots & \vdots & a_{ii}^{(i-1)} & \cdots & \cdots & a_{in}^{(i-1)} \\ \vdots & \vdots & 0 & a_{i+1,i+1}^{(i)} & \cdots & a_{i+1,n}^{(i)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{n,i+1}^{(i)} & \cdots & a_{nn}^{(i)} \end{bmatrix} \end{aligned} \tag{2.2.2}$$

If $a_{ii}^{(i-1)} \neq 0$, for $i = 1, \ldots, n-1$, then the method is executable and we have that

$$A^{(n-1)} = L_{n-1} \ldots L_1 A^{(0)} = R \tag{2.2.3}$$

is an upper triangular matrix. Thus, $A = LR$. Explicit representation of L:

$$\begin{aligned} L_i &= I - l_i e_i^T, \quad L_i^{-1} = I + l_i e_i^T \\ L &= L_1^{-1} \ldots L_{n-1}^{-1} = (I + l_1 e_1^T) \ldots (I + l_{n-1} e_{n-1}^T) \\ &= I + l_1 e_1^T + \ldots + l_{n-1} e_{n-1}^T \quad \text{(by (2.1.6))}. \end{aligned}$$

Theorem 2.2.1 *Let A be nonsingular. Then A has an LR-factorization (A=LR) if and only if $k_i := det(A_i) \neq 0$, where A_i is the leading principal matrix of A, i.e.,*

$$A_i = \begin{bmatrix} a_{11} & \ldots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \ldots & a_{ii} \end{bmatrix},$$

for $i = 1, \ldots, n-1$.

Proof: (Necessity "⇒"): Since $A = LR$, we have

$$\begin{bmatrix} a_{11} & \ldots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \ldots & a_{ii} \end{bmatrix} = \begin{bmatrix} l_{11} & & \\ \vdots & \ddots & O \\ l_{i1} & \ldots & l_{ii} \end{bmatrix} \begin{bmatrix} r_{11} & & r_{1i} \\ O & \ddots & \\ & & r_{ii} \end{bmatrix}.$$

From $\det(A) \neq 0$ follows that $\det(L) \neq 0$ and $\det(R) \neq 0$. Thus, $l_{jj} \neq 0$ and $r_{jj} \neq 0$, for $j = 1, \ldots, n$. Hence $k_i = l_{11} \ldots l_{ii} r_{11} \ldots r_{ii} \neq 0$.

(Sufficiency "$\Leftarrow$"): From (2.2.2) we have

$$A^{(0)} = (L_1^{-1} \ldots L_i^{-1}) A^{(i)}.$$

Consider the $(i+1)$-th leading principle determinant. From (2.2.3) we have

$$\begin{bmatrix} a_{11} & \ldots & a_{i,i+1} \\ \vdots & & \vdots \\ a_{i+1} & \ldots & a_{i+1,i+1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & & & & 0 \\ l_{21} & \ddots & & & \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ l_{i+1,1} & \cdots & \cdots & l_{i+1,i} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & \cdots & * \\ & a_{22}^{(1)} & \cdots & \cdots & \vdots \\ & & \ddots & & \vdots \\ & & & a_{ii}^{(i-1)} & a_{i,i+1}^{(i-1)} \\ 0 & & & & a_{i+1,i+1}^{(i)} \end{bmatrix}.$$

Thus, $k_i = 1 \cdot a_{11}^{(0)} a_{22}^{(1)} \ldots a_{i+1,i+1}^{(i)} \neq 0$ which implies $a_{i+1,i+1}^{(i)} \neq 0$. Therefore, the LR-factorization of A exists. ■

Theorem 2.2.2 *If a nonsingular matrix A has an LR-factorization with $A = LR$ and $l_{11} = \cdots = l_{nn} = 1$, then the factorization is unique.*

Proof: Let $A = L_1 R_1 = L_2 R_2$. Then $L_2^{-1} L_1 = R_2 R_1^{-1} = I$. ■

Corollary 2.2.1 *If a nonsingular matrix A has an LR-factorization with $A = LDR$, where D is diagonal, L and R^T are unit lower triangular (with one on the diagonal) if and only if $k_i \neq 0$.*

Theorem 2.2.3 *Let A be a nonsingular matrix. Then there exists a permutation P, such that PA has an LR-factorization.*

(Proof): By construction! Consider (2.2.2): There is a permutation P_i, which interchanges the i-th row with a row of index large than i, such that $0 \neq a_{ii}^{(i-1)} (\in P_i A^{(i-1)})$. This procedure is executable, for $i = 1, \ldots, n-1$. So we have

$$L_{n-1} P_{n-1} \ldots L_i P_i \ldots L_1 P_1 A^{(0)} = R. \tag{2.2.4}$$

Let P be a permutation which affects only elements $i+1, \cdots, n$. It holds

$$P(I - l_i e_i^T) P^{-1} = I - (P l_i) e_i^T = I - \tilde{l}_i e_i^T = \tilde{L}_i, \quad (e_i^T P^{-1} = e_i^T)$$

where $\tilde{L}_i$ is lower triangular. Hence we have

$$P L_i = \tilde{L}_i P. \tag{2.2.5}$$

Now write all P_i in (2.2.4) to the right as

$$L_{n-1} \tilde{L}_{n-2} \ldots \tilde{L}_1 P_{n-1} \ldots P_1 A^{(0)} = R.$$

Then we have $PA = LR$ with $L^{-1} = L_{n-1} \tilde{L}_{n-2} \cdots \tilde{L}_1$ and $P = P_{n-1} \cdots P_1$. ■

2.3 Gaussian elimination

2.3.1 Practical implementation

Given a linear system

$$Ax = b \tag{2.3.1}$$

with A nonsingular. We first assume that A has an LR-factorization. i.e., $A = LR$. Thus

$$LRx = b.$$

We then (i) solve $Ly = b$; (ii) solve $Rx = y$. These imply that $LRx = Ly = b$. From (2.2.4), we have

$$L_{n-1} \dots L_2 L_1 (A \mid b) = (R \mid L^{-1} b).$$

Algorithm 2.3.1 (without permutation)

For $k = 1, \dots, n-1$,
  if $a_{kk} = 0$ *then stop* $(*)$;
  else $\omega_j := a_{kj}$ $(j = k+1, \dots, n)$;
  for $i = k+1, \dots, n$,
    $\eta := a_{ik}/a_{kk}$, $a_{ik} := \eta$;
    for $j = k+1, \dots, n$,
      $a_{ij} := a_{ij} - \eta\omega_j$, $b_j := b_j - \eta b_k$.

For x: *(back substitution!)*
  $x_n = b_n/a_{nn}$;
  for $i = n-1, n-2, \dots, 1$,
    $x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$.

Cost of computation (one multiplication + one addition ≡ one flop):

(i) LR-factorization: $n^3/3 - n/3$ flops;

(ii) Computation of y: $n(n-1)/2$ flops;

(iii) Computation of x: $n(n+1)/2$ flops.

For A^{-1}: $4/3n^3 \approx n^3/3 + kn^2$ ($k = n$ linear systems).

Pivoting: (a) Partial pivoting; (b) Complete pivoting.

From (2.2.2), we have

$$A^{(k-1)} = \begin{bmatrix} a_{11}^{(0)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(0)} \\ 0 & \ddots & & & & \vdots \\ \vdots & & a_{k-1,k-1}^{(k-2)} & \cdots & \cdots & a_{k-1,n}^{(k-2)} \\ \vdots & & 0 & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{bmatrix}.$$

For (a):

$$\begin{cases} \text{Find a } \ p \in \{k, \ldots, n\} \text{ such that} \\ \qquad |a_{pk}| = \max_{k \le i \le n} |a_{ik}| \quad (r_k = p) \\ \text{swap } a_{kj}, b_k \text{ and } a_{pj}, b_p \text{ respectively, } (j = 1, \ldots, n). \end{cases} \tag{2.3.2}$$

Replacing $(*)$ in Algorithm 2.3.1 by (2.3.2), we have a new factorization of A with partial pivoting, i.e., $PA = LR$ (by Theorem 2.2.1) and $|l_{ij}| \le 1$ for $i, j = 1, \ldots, n$. For solving linear system $Ax = b$, we use

$$PAx = Pb \Rightarrow L(Rx) = P^T b \equiv \tilde{b}.$$

It needs extra $n(n-1)/2$ comparisons.

For (b):

$$\begin{cases} \text{Find } \ p, q \in \{k, \ldots, n\} \text{ such that} \\ \qquad |a_{pq}| \le \max\limits_{k \le i,j \le n} |a_{ij}|, \ (r_k := p, c_k := q) \\ \text{swap } \ a_{kj}, b_k \text{ and } a_{pj}, b_p \text{ respectively, } (j = k, \ldots, n), \\ \text{swap } \ a_{ik} \text{ and } a_{iq} (i = 1, \ldots, n). \end{cases} \tag{2.3.3}$$

Replacing $(*)$ in Algorithm 2.3.1 by (2.3.3), we also have a new factorization of A with complete pivoting, i.e., $PA\Pi = LR$ (by Theorem 2.2.1) and $|l_{ij}| \le 1$, for $i, j = 1, \ldots, n$. For solving linear system $Ax = b$, we use

$$PA\Pi(\Pi^T x) = Pb \Rightarrow LR\tilde{x} = \tilde{b} \Rightarrow x = \Pi\tilde{x}.$$

It needs $n^3/3$ comparisons.

Example 2.3.1 Let $A = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ be in three decimal-digit floating point arithmetic. Then $\kappa(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 4$. A is well-conditioned.

• Without pivoting:

$$\begin{aligned} L &= \begin{bmatrix} 1 & 0 \\ fl(1/10^{-4}) & 1 \end{bmatrix}, \quad fl(1/10^{-4}) = 10^4, \\ R &= \begin{bmatrix} 10^{-4} & 1 \\ 0 & fl(1 - 10^4 \cdot 1) \end{bmatrix}, \quad fl(1 - 10^4 \cdot 1) = -10^4. \\ LR &= \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix} \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix} \neq \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix} = A. \end{aligned}$$

Here a_{22} entirely "lost" from computation. It is numerically unstable. Let $Ax = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Then $x \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. But $Ly = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ solves $y_1 = 1$ and $y_2 = fl(2 - 10^4 \cdot 1) = -10^4$, $R\hat{x} = y$ solves $\hat{x}_2 = fl((-10^4)/(-10^4)) = 1$, $\hat{x}_1 = fl((1-1)/10^{-4}) = 0$. We have an erroneous solution with $\text{cond}(L)$, $\text{cond}(R) \approx 10^8$.

• Partial pivoting:

$$\begin{aligned} L &= \begin{bmatrix} 1 & 0 \\ fl(10^{-4}/1) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^{-4} & 1 \end{bmatrix}, \\ R &= \begin{bmatrix} 1 & 1 \\ 0 & fl(1 - 10^{-4}) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

L and R are both well-conditioned.

2.3.2 LDR- and LL^T-factorizations

Let $A = LDR$ as in Corollary 2.2.1.

Algorithm 2.3.2 (Crout's factorization or compact method)

For $k = 1, \ldots, n$,
  for $p = 1, 2, \ldots, k-1$,
    $r_p := d_p a_{pk}$,
    $\omega_p := a_{kp} d_p$,
    $d_k := a_{kk} - \sum_{p=1}^{k-1} a_{kp} r_p$,
  if $d_k = 0$, *then stop; else*
    for $i = k+1, \ldots, n$,
      $a_{ik} := (a_{ik} - \sum_{p=1}^{k-1} a_{ip} r_p)/d_k$,
      $a_{ki} := (a_{ki} - \sum_{p=1}^{k-1} \omega_p a_{pi})/d_k$.

Cost: $n^3/3$ flops.

- With partial pivoting: see Wilkinson EVP pp.225-.
- Advantage: One can use double precision for inner product.

Theorem 2.3.1 *If A is nonsingular, real and symmetric, then A has a unique LDL^T-factorization, where D is diagonal and L is a unit lower triangular matrix (with one on the diagonal).*

Proof: $A = LDR = A^T = R^T D L^T$. It implies $L = R^T$. ■

Theorem 2.3.2 *If A is symmetric and positive definite, then there exists a lower triangular $G \in \mathbb{R}^{n\times n}$ with positive diagonal elements such that $A = GG^T$.*

Proof: A is symmetric positive definite $\Leftrightarrow x^T A x \geq 0$, for all nonzero vector $x \in \mathbb{R}^{n\times n}$ $\Leftrightarrow k_i \geq 0$, for $i = 1, \cdots, n$, $\Leftrightarrow$ all eigenvalues of A are positive.

From Corollary 2.2.1 and Theorem 2.3.1 we have $A = LDL^T$. From $L^{-1}AL^{-T} = D$ follows that $d_k = (e_k^T L^{-1}) A (L^{-T} e_k) > 0$. Thus, $G = L\mathrm{diag}\{d_1^{1/2}, \cdots, d_n^{1/2}\}$ is real, and then $A = GG^T$. ■

Algorithm 2.3.3 (Cholesky factorization) *Let A be symmetric positive definite. To find a lower triangular matrix G such that $A = GG^T$.*

For $k = 1, 2, \ldots, n$,
  $a_{kk} := (a_{kk} - \sum_{p=1}^{k-1} a_{kp}^2)^{1/2}$;
  for $i = k+1, \ldots, n$,
    $a_{ik} = (a_{ik} - \sum_{p=1}^{k-1} a_{ip} a_{kp})/a_{kk}$.

Cost: $n^3/6$ flops.

Remark 2.3.1 *For solving symmetric, indefinite systems: See Golub/ Van Loan Matrix Computation pp. 159-168. $PAP^T = LDL^T$, D is 1×1 or 2×2 block-diagonal matrix, P is a permutation and L is lower triangular with one on the diagonal.*

2.3.3 Error estimation for linear systems

Consider the linear system

$$Ax = b, \tag{2.3.4}$$

and the perturbed linear system

$$(A + \delta A)(x + \delta x) = b + \delta b, \tag{2.3.5}$$

where δA and δb are errors of measure or round-off in factorization.

Definition 2.3.1 *Let $\|\quad\|$ be an operator norm and A be nonsingular. Then $\kappa \equiv \kappa(A) = \|A\|\|A^{-1}\|$ is a condition number of A corresponding to $\|\ \|$.*

Theorem 2.3.3 (Forward error bound) *Let x be the solution of the (2.3.4) and $x+\delta x$ be the solution of the perturbed linear system (2.3.5). If $\|\delta A\|\|A^{-1}\| < 1$, then*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa}{1 - \kappa\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \tag{2.3.6}$$

Proof: From (2.3.5) we have

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

Thus,

$$\delta x = -(A + \delta A)^{-1}[(\delta A)x - \delta b]. \tag{2.3.7}$$

Here, Corollary 2.7 implies that $(A + \delta A)^{-1}$ exists. Now,

$$\|(A + \delta A)^{-1}\| = \|(I + A^{-1}\delta A)^{-1}A^{-1}\| \leq \|A^{-1}\|\frac{1}{1 - \|A^{-1}\|\|\delta A\|}.$$

On the other hand, $b = Ax$ implies $\|b\| \leq \|A\|\|x\|$. So,

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{2.3.8}$$

From (2.3.7) follows that $\|\delta x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta A\|\|x\| + \|\delta b\|)$. By using (2.3.8), the inequality (2.3.6) is proved. ■

Remark 2.3.2 *If $\kappa(A)$ is large, then A (for the linear system $Ax = b$) is called ill-conditioned, else well-conditioned.*

2.3.4 Error analysis for Gaussian algorithm

A computer in characterized by four integers: (a) the machine base β; (b) the precision t; (c) the underflow limit L; (d) the overflow limit U. Define the set of floating point numbers.

$$F = \{f = \pm 0.d_1d_2\cdots d_t \times \beta^e \mid 0 \leq d_i < \beta, d_1 \neq 0, L \leq e \leq U\} \cup \{0\}. \tag{2.3.9}$$

Let $G = \{x \in \mathbb{R} \mid m \le |x| \le M\} \cup \{0\}$, where $m = \beta^{L-1}$ and $M = \beta^U(1-\beta^{-t})$ are the minimal and maximal numbers of $F \setminus \{0\}$ in absolute value, respectively. We define an operator $fl : G \to F$ by

$$fl(x) = \text{ the nearest } c \in F \text{ to } x \text{ by rounding arithmetic.}$$

One can show that fl satisfies

$$fl(x) = x(1+\varepsilon), \quad |\varepsilon| \le eps, \tag{2.3.10}$$

where $eps = \frac{1}{2}\beta^{1-t}$. (If $\beta = 2$, then $eps = 2^{-t}$). It follows that

$$fl(a \circ b) = (a \circ b)(1+\varepsilon)$$

or

$$fl(a \circ b) = (a \circ b)/(1+\varepsilon),$$

where $|\varepsilon| \le eps$ and $\circ = +, -, \times, /$.

Algorithm 2.3.4 *Given $x, y \in \mathbb{R}^n$. The following algorithm computes $x^T y$ and stores the result in s.*

$s = 0$,
for $k = 1, \ldots, n$,
$\quad s = s + x_k y_k$.

Theorem 2.3.4 *If $n2^{-t} \le 0.01$, then*

$$fl(\sum_{k=1}^{n} x_k y_k) = \sum_{k=1}^{n} x_k y_k[1 + 1.01(n+2-k)\theta_k 2^{-t}], \ |\theta_k| \le 1$$

Proof: Let $s_p = fl(\sum_{k=1}^{p} x_k y_k)$ be the partial sum in Algorithm 2.3.4. Then

$$s_1 = x_1 y_1(1+\delta_1)$$

with $|\delta_1| \le eps$ and for $p = 2, \ldots, n$,

$$s_p = fl[s_{p-1} + fl(x_p y_p)] = [s_{p-1} + x_p y_p(1+\delta_p)](1+\varepsilon_p)$$

with $|\delta_p|$, $|\varepsilon_p| \le eps$. Therefore

$$fl(x^T y) = s_n = \sum_{k=1}^{n} x_k y_k(1+\gamma_k),$$

where $(1+\gamma_k) = (1+\delta_k)\prod_{j=k}^{n}(1+\varepsilon_j)$, and $\varepsilon_1 \equiv 0$. Thus,

$$fl(\sum_{k=1}^{n} x_k y_k) = \sum_{k=1}^{n} x_k y_k[1 + 1.01(n+2-k)\theta_k 2^{-t}]. \tag{2.3.11}$$

The result follows immediately from the following useful Lemma. ■

Lemma 2.3.5 *If* $(1+\alpha) = \prod_{k=1}^{n}(1+\alpha_k)$, *where* $|\alpha_k| \le 2^{-t}$ *and* $n2^{-t} \le 0.01$, *then*

$$\prod_{k=1}^{n}(1+\alpha_k) = 1 + 1.01n\theta 2^{-t} \text{ with } |\theta| \le 1.$$

Proof: From assumption it is easily seen that

$$(1-2^{-t})^n \le \prod_{k=1}^{n}(1+\alpha_k) \le (1+2^{-t})^n. \tag{2.3.12}$$

Expanding the Taylor expression of $(1-x)^n$ as $-1 < x < 1$, we get

$$(1-x)^n = 1 - nx + \frac{n(n-1)}{2}(1-\theta x)^{n-2}x^2 \ge 1 - nx.$$

Hence

$$(1-2^{-t})^n \ge 1 - n2^{-t}. \tag{2.3.13}$$

Now, we estimate the upper bound of $(1+2^{-t})^n$:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = 1 + x + \frac{x}{2}x(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots).$$

If $0 \le x \le 0.01$, then

$$1 + x \le e^x \le 1 + x + 0.01x\frac{1}{2}e^x \le 1 + 1.01x \tag{2.3.14}$$

(Here, we use the fact $e^{0.01} < 2$ to the last inequality.) Let $x = 2^{-t}$. Then the left inequality of (2.3.14) implies

$$(1+2^{-t})^n \le e^{2^{-t}n} \tag{2.3.15}$$

Let $x = 2^{-t}n$. Then the second inequality of (2.3.14) implies

$$e^{2^{-t}n} \le 1 + 1.01n2^{-t} \tag{2.3.16}$$

From (2.3.15) and (2.3.16) we have

$$(1+2^{-t})^n \le 1 + 1.01n2^{-t}.$$

■

Let the exact LR-factorization of A be L and R ($A = LR$) and let $\tilde{L}$, $\tilde{R}$ be the LR-factorization of A by using Gaussian Algorithm (without pivoting). There are two possibilities:

(i) Forward error analysis: Estimate $|L - \tilde{L}|$ and $|R - \tilde{R}|$.

(ii) Backward error analysis: Let $\tilde{L}\tilde{R}$ be the exact LR-factorization of a perturbed matrix $\tilde{A} = A + F$. Then F will be estimated, i.e., $|F| \le ?$.

2.3.5 Àpriori error estimate for backward error bound of LR-factorization

From (2.2.2) we have

$$A^{(k+1)} = L_k A^{(k)},$$

for $k = 1, 2, \ldots, n-1$ ($A^{(1)} = A$). Denote the entries of $A^{(k)}$ by $a_{ij}^{(k)}$ and let $l_{ik} = fl(a_{ik}^{(k)}/a_{kk}^{(k)})$, $i \geq k+1$. From (2.2.2) we know that

$$a_{ij}^{(k+1)} = \begin{cases} 0; & \text{for } i \geq k+1,\ j = k \\ fl(a_{ij}^{(k)} - fl(l_{ik}a_{kj}^{(k)})); & \text{for } i \geq k+1,\ j \geq k+1 \\ a_{ij}^{(k)}; & \text{otherwise.} \end{cases} \tag{2.3.17}$$

From (2.3.10) we have $l_{ik} = (a_{ik}^{(k)}/a_{kk}^{(k)})(1+\delta_{ik})$ with $|\delta_{ik}| \leq 2^{-t}$. Then

$$a_{ik}^{(k)} - l_{ik}a_{kk}^{(k)} + a_{ij}^{(k)}\delta_{ik} = 0, \quad \text{for } i \geq k+1. \tag{2.3.18}$$

Let $a_{ik}^{(k)}\delta_{ik} \equiv \varepsilon_{ik}^{(k)}$. From (2.3.10) we also have

$$\begin{aligned} a_{ij}^{(k+1)} &= fl(a_{ij}^{(k)} - fl(l_{ik}a_{kj}^{(k)})) \\ &= (a_{ij}^{(k)} - (l_{ik}a_{kj}^{(k)}(1+\delta_{ij})))/(1+\delta_{ij}^{'}) \end{aligned} \tag{2.3.19}$$

with $|\delta_{ij}|$, $|\delta_{ij}^{'}| \leq 2^{-t}$. Then

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} - l_{ik}a_{kj}^{(k)}\delta_{ij} + a_{ij}^{(k+1)}\delta_{ij}^{'}, \quad \text{for } i,j \geq k+1. \tag{2.3.20}$$

Let $\varepsilon_{ij}^{(k)} \equiv -l_{ik}a_{kj}^{(k)}\delta_{ij} + a_{ij}^{(k+1)}\delta_{ij}^{'}$ which is the computational error of $a_{ij}^{(k)}$ in $A^{(k+1)}$. From (2.3.17), (2.3.18) and (2.3.20) we obtain

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} - l_{ik}a_{kk}^{(k)} + \varepsilon_{ij}^{(k)}; & \text{for } i \geq k+1, j = k \\ a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} + \varepsilon_{ij}^{(k)}; & \text{for } i \geq k+1, j \geq k+1 \\ a_{ij}^{(k)} + \varepsilon_{ij}^{(k)}; & \text{otherwise,} \end{cases} \tag{2.3.21}$$

where

$$\varepsilon_{ij}^{(k)} = \begin{cases} a_{ij}^{(k)}\delta_{ij}; & \text{for } i \geq k+1, j = k, \\ -l_{ik}a_{kj}^{(k)}\delta_{ij} - a_{ij}^{(k+1)}\delta_{ij}^{'}; & \text{for } i \geq k+1, j \geq k+1 \\ 0; & \text{otherwise.} \end{cases} \tag{2.3.22}$$

Let $E^{(k)}$ be the error matrix with entries $\varepsilon_{ij}^{(k)}$. Then (2.3.21) can be written as

$$A^{(k+1)} = A^{(k)} - M_k A^{(k)} + E^{(k)}, \tag{2.3.23}$$

where

$$M_k = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & l_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & l_{n,k} & & 0 \end{bmatrix} \tag{2.3.24}$$

For $k = 1, 2 \ldots, n-1$, we add the $n-1$ equations in (2.3.23) together and get

$$\begin{aligned} M_1 A^{(1)} &+ M_2 A^{(2)} + \cdots + M_{n-1} A^{(n-1)} + A^{(n)} \\ &= A^{(1)} + E^{(1)} + \cdots + E^{(n-1)}. \end{aligned}$$

From (2.3.17) we know that the k-th row of $A^{(k)}$ is equal to the k-th row of $A^{(k+1)}, \cdots, A^{(n)}$, respectively and from (2.3.24) we also have

$$M_k A^{(k)} = M_k A^{(n)} = M_k \tilde{R}.$$

Thus,

$$(M_1 + M_2 + \cdots + M_{n-1} + I)\tilde{R} = A^{(1)} + E^{(1)} + \cdots + E^{(n-1)}.$$

Then

$$\tilde{L}\tilde{R} = A + E, \tag{2.3.25}$$

where

$$\tilde{L} = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & O \\ \vdots & & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix} \text{ and } E = E^{(1)} + \cdots + E^{(n-1)}. \tag{2.3.26}$$

Now we assume that the partial pivotings in Gaussian Elimination are already arranged such that pivot element $a_{kk}^{(k)}$ has the maximal absolute value. So, we have $|l_{ik}| \leq 1$. Let

$$\rho = \max_{i,j,k} |a_{ij}^{(k)}| / \|A\|_\infty. \tag{2.3.27}$$

Then

$$|a_{ij}^{(k)}| \leq \rho \|A\|_\infty. \tag{2.3.28}$$

From (2.3.22) and (2.3.28) follows that

$$|\varepsilon_{ij}^{(k)}| \leq \rho \|A\|_\infty \begin{cases} 2^{-t}; & \text{for } i \geq k+1,\ j = k, \\ 2^{1-t}; & \text{for } i \geq k+1,\ j \geq k+1, \\ 0; & \text{otherwise.} \end{cases} \tag{2.3.29}$$

Therefore,

$$|E^{(k)}| \leq \rho \|A\|_\infty 2^{-t} \cdot \left[\begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & \cdots & 2 \end{array}\right]. \tag{2.3.30}$$

From (2.3.26) we get

$$|E| \leq \rho \|A\|_\infty \cdot 2^{-t} \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 2 & 2 & \cdots & 2 & 2 \\ 1 & 3 & 4 & \cdots & 4 & 4 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 3 & 5 & \cdots & 2n-4 & 2n-4 \\ 1 & 3 & 5 & \cdots & 2n-3 & 2n-2 \end{bmatrix} \tag{2.3.31}$$

Hence we have the following theorem.

Theorem 2.3.6 *The LR-factorization $\tilde{L}$ and $\tilde{R}$ of A using Gaussian Elimination with partial pivoting satisfies*

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

where

$$\|E\|_\infty \leq n^2 \rho \|A\|_\infty 2^{-t} \tag{2.3.32}$$

Proof:

$$\|E\|_\infty \leq \rho \|A\|_\infty 2^{-t} (\sum_{j=1}^{n} (2j-1) - 1) < n^2 \rho \|A\|_\infty 2^{-t}.$$

■

Now we shall solve the linear system $Ax = b$ by using the factorization $\tilde{L}$ and $\tilde{R}$, i.e., $\tilde{L}y = b$ and $\tilde{R}x = y$.

• For $Ly = b$: From Algorithm 2.3.1 we have

$$\begin{aligned} y_1 &= fl(b_1/l_{11}), \\ y_i &= fl\left(\frac{-l_{i1}y_1 - l_{i2}y_2 - \cdots - l_{i,i-1}y_{i-1} + b_i}{l_{ii}}\right), \end{aligned} \tag{2.3.33}$$

for $i = 2, 3, \ldots, n$. From (2.3.10) we have

$$\begin{cases} y_1 = b_1/l_{11}(1+\delta_{11}), \text{ with } |\delta_{11}| \leq 2^{-t} \\ y_i = fl(\frac{fl(-l_{i1}y_1 - l_{i2}y_2 - \cdots - l_{i,i-1}y_{i-1}) + b_i}{l_{ii}(1+\delta_{ii})}) \\ \quad = \frac{fl(-l_{i1}y_1 - l_{i2}y_2 - \cdots - l_{i,i-1}y_{i-1}) + b_i}{l_{ii}(1+\delta_{ii})(1+\delta'_{ii})}, \text{ with } |\delta_{ii}|,\ |\delta'_{ii}| \leq 2^{-t}. \end{cases} \tag{2.3.34}$$

Applying Theorem 2.3.4 we get

$$fl(-l_{i1}y_1 - l_{i2}y_2 - \cdots - l_{i,i-1}y_{i-1}) = -l_{i1}(1+\delta_{i1})y_1 - \cdots - l_{i,i-1}(1+\delta_{i,i-1})y_{i-1},$$

where

$$\begin{aligned} |\delta_{i1}| &\leq (i-1)1.01 \cdot 2^{-t};\ \text{for } i = 2, 3, \cdots, n, \\ |\delta_{ij}| &\leq (i+1-j)1.01 \cdot 2^{-t};\ \text{for } \begin{cases} i = 2, 3, \cdots, n, \\ j = 2, 3, \cdots, i-1. \end{cases} \end{aligned} \tag{2.3.35}$$

So, (2.3.34) can be written as

$$\begin{cases} l_{11}(1+\delta_{11})y_1 = b_1, \\ l_{i1}(1+\delta_{i1})y_1 + \cdots + l_{i,i-1}(1+\delta_{i,i-1})y_{i-1} + l_{ii}(1+\delta_{ii})(1+\delta'_{ii})y_i = b_i, \\ \text{for } i = 2, 3, \cdots, n. \end{cases} \tag{2.3.36}$$

or

$$(L + \delta L)y = b. \tag{2.3.37}$$

From (2.3.35) (2.3.36) and (2.3.37) follow that

$$|\delta L| \leq 1.01 \cdot 2^{-t} \begin{bmatrix} |l_{11}| & & & & & 0 \\ |l_{21}| & 2|l_{22}| & & & & \\ 2|l_{31}| & 2|l_{32}| & 2|l_{33}| & & & \\ 3|l_{41}| & 3|l_{42}| & 2|l_{43}| & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ (n-1)|l_{n1}| & (n-1)|l_{n2}| & (n-2)|l_{n3}| & \cdots & 2|l_{n,n-1}| & 2|l_{nn}| \end{bmatrix}. \tag{2.3.38}$$

This implies,

$$\|\delta L\|_\infty \leq \frac{n(n+1)}{2} \cdot 1.01 \cdot 2^{-t} \max_{i,j} |l_{ij}| \leq \frac{n(n+1)}{2} \cdot 1.01 \cdot 2^{-t}. \tag{2.3.39}$$

Theorem 2.3.7 *For lower triangular linear system $Ly = b$, if y is the exact solution of $(L + \delta L)y = b$, then δL satisfies (2.3.38) and (2.3.39).* ■

Applying Theorem 2.3.7 to the linear system $\tilde{L}y = b$ and $\tilde{R}x = y$, respectively, the solution x satisfies

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

or

$$(\tilde{L}\tilde{R} + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R}))x = b. \tag{2.3.40}$$

Since $\tilde{L}\tilde{R} = A + E$, substituting this equation into (2.3.40) we get

$$[A + E + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R})]x = b. \tag{2.3.41}$$

The entries of $\tilde{L}$ and $\tilde{R}$ satisfy

$$|\widetilde{l_{ij}}| \leq 1, \text{ and } |\widetilde{r}_{ij}| \leq \rho\|A\|_\infty.$$

Therefore, we get

$$\begin{cases} \|\tilde{L}\|_\infty \leq n, \\ \|\tilde{R}\|_\infty \leq n\rho\|A\|_\infty, \\ \|\delta\tilde{L}\|_\infty \leq \frac{n(n+1)}{2} 1.01 \cdot 2^{-t}, \\ \|\delta\tilde{R}\|_\infty \leq \frac{n(n+1)}{2} 1.01\rho 2^{-t}. \end{cases} \tag{2.3.42}$$

In practical implementation we usually have $n^2 2^{-t} << 1$. So it holds

$$\|\delta\tilde{L}\|_\infty \|\delta\tilde{R}\|_\infty \leq n^2\rho\|A\|_\infty 2^{-t}.$$

Let

$$\delta A = E + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R}). \tag{2.3.43}$$

Then, (2.3.32) and (2.3.42) we get

$$\begin{aligned} \|\delta A\|_\infty &\leq \|E\|_\infty + \|\delta\tilde{L}\|_\infty\|\tilde{R}\|_\infty + \|\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty + \|\delta\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty \\ &\leq 1.01(n^3 + 3n^2)\rho\|A\|_\infty 2^{-t} \end{aligned} \tag{2.3.44}$$

Theorem 2.3.8 *For a linear system $Ax = b$ the solution x computed by Gaussian Elimination with partial pivoting is the exact solution of the equation $(A + \delta A)x = b$ and δA satisfies (2.3.43) and (2.3.44).*

Remark 2.3.3 The quantity ρ defined by (2.3.27) is called a growth factor. The growth factor measures how large the numbers become during the process of elimination. In practice, ρ is usually of order 10 for partial pivot selection. But it can be as large as $\rho = 2^{n-1}$, when

$$A = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ \vdots & -1 & \ddots & \ddots & \vdots & 1 \\ \vdots & \vdots & \ddots & \ddots & 0 & 1 \\ -1 & -1 & \cdots & -1 & 1 & 1 \\ -1 & -1 & \cdots & \cdots & -1 & 1 \end{bmatrix}.$$

Better estimates hold for special types of matrices. For example in the case of upper Hessenberg matrices, that is, matrices of the form

$$A = \begin{bmatrix} \times & \cdots & \cdots & \times \\ \times & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & \times & \times \end{bmatrix}$$

the bound $\rho \leq (n-1)$ can be shown. (Hessenberg matrices arise in eigenvalus problems.)

For tridiagonal matrices

$$A = \begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \gamma_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_n \\ 0 & & & \gamma_n & \alpha_n \end{bmatrix}$$

it can even be shown that $\rho \leq 2$ holds for partial pivot selection. Hence, Gaussian elimination is quite numerically stable in this case.

For complete pivot selection, Wilkinson (1965) has shown that

$$|a_{ij}^k| \leq f(k) \max_{i,j} |a_{ij}|$$

with the function

$$f(k) := k^{\frac{1}{2}}[2^1 3^{\frac{1}{2}} 4^{\frac{1}{3}} \cdots k^{\frac{1}{(k-1)}}]^{\frac{1}{2}}.$$

This function grows relatively slowly with k:

k	10	20	50	100
$f(k)$	19	67	530	3300

Even this estimate is too pessimistic in practice. Up until now, no matrix has been found which fails to satisfy

$$|a_{ij}^{(k)}| \leq (k+1)\max_{i,j}|a_{ij}| \quad k = 1, 2, ..., n-1,$$

when complete pivot selection is used. This indicates that Gaussian elimination with complete pivot selection is usually a stable process. Despite this, partial pivot selection is preferred in practice, for the most part, because:

(i) Complete pivot selection is more costly than partial pivot selection. (To compute $A^{(i)}$, the maximum from among $(n-i+1)^2$ elements must be determined instead of $n-i+1$ elements as in partial pivot selection.)

(ii) Special structures in a matrix, i.e. the band structure of a tridiagonal matrix, are destroyed in complete pivot selection.

2.3.6 Improving and Estimating Accuracy

• Iterarive Improvement:

Suppose that the linear system $Ax = b$ has been solved via the LR-factorization $PA = LR$. Now we want to improve the accuracy of the computed solution x. We compute

$$\begin{cases} r &= b - Ax, \\ Ly &= Pr, \quad Rz = y, \\ x_{new} &= x + z. \end{cases} \tag{2.3.45}$$

Then in exact arithmatic we have

$$Ax_{new} = A(x+z) = (b-r) + Az = b.$$

Unfortunately, $r = fl(b - Ax)$ renders an x_{new} that is no more accurate than x. It is necessary to compute the residual $b - Ax$ with extended precision floating arithmetic.

Algorithm 2.3.5

Compute $PA = LR$ *(t-digit)*
Repeat: $r := b - Ax$ *(2t-digit)*
  Solve $Ly = Pr$ *for* y *(t-digit)*
  Solve $Rz = y$ *for* z *(t-digit)*
  Update $x = x + z$ *(t-digit)*

This is referred to as an iterative improvement. From (2.3.45) we have

$$r_i = b_i - a_{i1}x_1 - a_{i2}x_2 - \cdots - a_{in}x_n. \tag{2.3.46}$$

Now, r_i can be roughly estimated by $2^{-t}\max_j |a_{ij}|\,|x_j|$. That is

$$\|r\| \approx 2^{-t}\|A\|\|x\|. \tag{2.3.47}$$

Let $e = x - A^{-1}b = A^{-1}(Ax - b) = -A^{-1}r$. Then we have

$$\|e\| \leq \|A^{-1}\|\|r\|. \tag{2.3.48}$$

From (2.3.47) follows that

$$\|e\| \approx \|A^{-1}\| \cdot 2^{-t}\|A\|\|x\| = 2^{-t}\text{cond}(A)\|x\|.$$

Let

$$\text{cond}(A) = 2^p,\ 0 < p < t,\ (p \text{ is integer}). \tag{2.3.49}$$

Then we have

$$\|e\|/\|x\| \approx 2^{-(t-p)}. \tag{2.3.50}$$

From (2.3.50) we know that x has $q = t-p$ correct significant digits. Since r is computed by double precision, so we can assume that it has at least t correct significant digits. Therefore for solving $Az = r$ according to (2.3.50) the solution z (comparing with $-e = A^{-1}r$) has q-digits accuracy so that $x_{new} = x + z$ has usually $2q$-digits accuracy. From above discussion, the accuracy of x_{new} is improved about q-digits after one iteration. Hence we stop the iteration, when the number of the iterates k (say!) satifies $kq \geq t$. From above we have

$$\|z\|/\|x\| \approx \|e\|/\|x\| \approx 2^{-q} = 2^{-t}2^p. \tag{2.3.51}$$

From (2.3.49) and (2.3.51) we have

$$\text{cond}(A) = 2^t \cdot (\|z\|/\|x\|).$$

By (2.3.51) we get

$$q = \log_2(\frac{\|x\|}{\|z\|}) \text{ and } k = \frac{t}{\log_2(\frac{\|x\|}{\|z\|})}.$$

In the following we shall give a further discussion of convergence of the iterative improvement. From Theorem 2.3.8 we know that z in Algorithm 5.5 is computed by $(A+\delta A)z = r$. That is

$$A(I + F)z = r, \tag{2.3.52}$$

where $F = A^{-1}\delta A$.

Theorem 2.3.9 *Let the sequence of vectors $\{x_v\}$ be the sequence of improved solutions in Algorithm 5.5 for solving $Ax = b$ and $x^* = A^{-1}b$ be the exact solution. Assume that F_k in (2.3.52) satisfies $\|F_k\| \leq \sigma < 1/2$ for all k. Then $\{x_k\}$ converges to x^*, i.e., $\lim_{v\to\infty}\|x_k - x^*\| = 0$.*

Proof: From (2.3.52) and $r_k = b - Ax_k$ we have

$$A(I + F_k)z_k = b - Ax_k. \tag{2.3.53}$$

Since A is nonsingular, multiplying both sides of (2.3.53) by A^{-1} we get

$$(I + F_k)z_k = x^* - x_k.$$

From $x_{k+1} = x_k + z_k$ we have $(I + F_k)(x_{k+1} - x_k) = x^* - x_k$, i.e.,

$$(I + F_k)x_{k+1} = F_k x_k + x^*. \tag{2.3.54}$$

Subtracting both sides of (2.3.54) from $(I + F_k)x^*$ we get

$$(I + F_k)(x_{k+1} - x^*) = F_k(x_k - x^*).$$

Applying Corollary 1.2.1 we have

$$x_{k+1} - x^* = (I + F_k)^{-1}F_k(x_k - x^*).$$

Hence,

$$\|x_{k+1} - x^*\| \leq \|F_k\| \frac{\|x_k - x^*\|}{1 - \|F_k\|} \leq \frac{\sigma}{1 - \sigma}\|x_k - x^*\|.$$

Let $\tau = \sigma/(1 - \sigma)$. Then

$$\|x_k - x^*\| \leq \tau^{k-1}\|x_1 - x^*\|.$$

But $\sigma < 1/2$ follows $\tau < 1$. This implies convergence of Algorithm 2.3.5. ■

Corollary 2.3.1 *If*

$$1.01(n^3 + 3n^2)\rho 2^{-t}\|A\|\,\|A^{-1}\| < 1/2,$$

then Algorithm 2.3.5 converges.

Proof: From (2.3.52) and (2.3.44) follows that

$$\|F_k\| \leq 1.01(n^3 + 3n^2)\rho 2^{-t}\text{cond}(A) < 1/2.$$

■

2.4 Special Linear Systems

2.4.1 Toeplitz Systems

Definition 2.4.1 **(i)** *$T \in \mathbb{R}^{n\times n}$ is called a Toeplitz matrix if there exists $r_{-n+1}, \cdots, r_0, \cdots, r_{n-1}$ such that $a_{ij} = r_{j-i}$ for all i, j. e.g.,*

$$T = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 \\ r_{-1} & r_0 & r_1 & r_2 \\ r_{-2} & r_{-1} & r_0 & r_1 \\ r_{-3} & r_{-2} & r_{-1} & r_0 \end{bmatrix}, \quad (n = 4).$$

(ii) *$B \in \mathbb{R}^{n\times n}$ is called a Persymmetric matrix if it is symmetric about northest-southwest diagonal, i.e., $b_{ij} = b_{n-j+1,n-i+1}$ for all i, j. That is,*

$$B = EB^T E, \text{ where } E = [e_n, \cdots e_1].$$

Given scalars $r_1, \cdots, r_{n-1}$ such that the matrices

$$T_k = \begin{bmatrix} 1 & r_1 & r_2 & \cdots & r_{k-1} \\ r_1 & 1 & r_1 & & \vdots \\ \vdots & & \ddots & & \\ r_{k-1} & \cdots & \cdots & & 1 \end{bmatrix}$$

are all positive definite, for $k = 1, \ldots, n$. Three algorithms will be described:

(a) Durbin's Algorithm for the Yule-Walker problem

$$T_n y = -(r_1, \ldots, r_n)^T.$$

(b) Levinson's Algorithm for the general right hand side $T_n x = b$.

(c) Trench's Algorithm for computing $B = T_n^{-1}$.

• To (a): Let $E_k = [e_k^{(k)}, \ldots, e_1^{(k)}]$. Suppose the k-th order Yule-Walker system

$$T_k y = -(r_1, \ldots, r_k)^T = -r^T$$

has been solved. Consider the $(k+1)$-st order system

$$\begin{bmatrix} T_k & E_k r \\ r^T E_k & 1 \end{bmatrix} \begin{bmatrix} z \\ \alpha \end{bmatrix} = \begin{bmatrix} -r \\ -r_{k+1} \end{bmatrix}$$

can be solved in $O(k)$ flops. Observe that

$$z = T_k^{-1}(-r - \alpha E_k r) = y - \alpha T_k^{-1} E_k r \tag{2.4.55}$$

and

$$\alpha = -r_{k+1} = -r^T E_k z. \tag{2.4.56}$$

Since T_k^{-1} is persymmetric, $T_k^{-1} E_k = E_k T_k^{-1}$ and $z = y + \alpha E_k y$. Substituting into (2.4.56) we get

$$\alpha = -r_{k+1} - r^T E_k(y + \alpha E_k y) = -(r_{k+1} + r^T E_k y)/(1 + r^T y).$$

Here $(1 + r^T y)$ is positive, because T_{k+1} is positive definite and

$$\begin{bmatrix} I & E_k y \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} T_k & E_k r \\ r^T E_k & 1 \end{bmatrix} \begin{bmatrix} I & E_k y \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} T_k & 0 \\ 0 & 1 + r^T y \end{bmatrix}.$$

Algorithm 2.4.1 (Durbin Algorithm, 1960) *Let* $T_k y^{(k)} = -r^{(k)} = -(r_1, \ldots, r_k)^T$.

For $k = 1, \ldots, n$,
$y^{(1)} = -r_1$,
for $k = 1, \ldots, n-1$,
$\beta_k = 1 + r^{(k)T} y^{(k)}$,
$\alpha_k = -(r_{k+1} + r^{(k)T} E_k y^{(k)})/\beta_k$,
$z^{(k)} = y^{(k)} + \alpha_k E_k y^{(k)}$,
$y^{(k+1)} = \begin{bmatrix} z^{(k)} \\ \alpha_k \end{bmatrix}$.

This algorithm requires $\frac{3}{2}n^2$ flops to generate $y = y^{(n)}$.
Further reduction:

$$\begin{aligned}\beta_k &= 1 + r^{(k)T}y^{(k)} \\ &= 1 + [r^{(k-1)T}, r^{(k)}] \begin{bmatrix} y^{(k-1)} + \alpha_{k-1}E_{k-1}y^{(k-1)} \\ \alpha_{k-1} \end{bmatrix} \\ &= 1 + r^{(k-1)T}y^{(k-1)} + \alpha_{k-1}(r^{(k-1)T}E_{k-1}y^{(k-1)} + r_k) \\ &= \beta_{k-1} + \alpha_{k-1}(-\beta_{k-1}\alpha_{k-1}) = (1 - \alpha_{k-1}^2)\beta_{k-1}.\end{aligned}$$

• To (b):

$$T_k x = b = (b_1, \cdots, b_k)^T, \text{ for } 1 \le k \le n. \tag{2.4.57}$$

Want to solve

$$\begin{bmatrix} T_k & E_k r \\ r^T E_k & 1 \end{bmatrix} \begin{bmatrix} \nu \\ \mu \end{bmatrix} \begin{bmatrix} b \\ b_{k+1} \end{bmatrix}, \tag{2.4.58}$$

where $r = (r_1, \cdots, r_k)^T$. Since $\nu = T_k^{-1}(b - \mu E_k r) = x + \mu E_k y$, it follows that

$$\begin{aligned}\mu &= b_{k+1} - r^T E_k \nu = b_{k+1} - r^T E_k x - \mu r^T y \\ &= (b_{k+1} - r^T E_k x)/(1 + r^T y).\end{aligned}$$

We can effect the transition form (2.4.57) to (2.4.58) in $O(k)$ flops. We can solve the system $T_n x = b$ by solving

$$T_k x^{(k)} = b^{(k)} = (b_1, \ldots, b_k)^T$$

and

$$T_k y^{(k)} = -r^{(k)} = -(r_1, \ldots, r_k)^T.$$

It needs $2n^2$ flops. See Algorithm Levinson (1947) in Matrix Computations, pp.128-129 for details.

• To (c):

$$T_n^{-1} = \begin{bmatrix} A & Er \\ r^T E & 1 \end{bmatrix}^{-1} = \begin{bmatrix} B & \nu \\ \nu^T & \gamma \end{bmatrix},$$

where $A = T_{n-1}, E = E_{n-1}$ and $r = (r_1, \ldots, r_{n-1})^T$. From the equation

$$\begin{bmatrix} A & Er \\ r^T E & 1 \end{bmatrix} = \begin{bmatrix} \nu \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

follows that

$$A\nu = -\gamma Er = -\gamma E(r_1, \ldots, r_{n-1})^T \text{ and } \gamma = 1 - r^T E\nu.$$

If y is the solution of $(n-1)$-st Yule-Walker system $Ay = -r$, then

$$\gamma = 1/(1 + r^T y) \text{ and } \nu = \gamma E y.$$

Thus the last row and column of T_n^{-1} are readily obtained. Since $AB + Er\nu^T = I_{n-1}$, we have

$$B = A^{-1} - (A^{-1}Er)\nu^T = A^{-1} + \frac{\nu\nu^T}{\gamma}.$$

Since $A = T_{n-1}$ is nonsingular and Toeplitz, its inverse is persymmetric. Thus

$$\begin{aligned} b_{ij} &= (A^{-1})_{ij} + \frac{\nu_i \nu_j}{\gamma} = (A^{-1})_{n-j,n-i} + \frac{\nu_i \nu_j}{\gamma} \\ &= b_{n-j,n-i} - \frac{\nu_{n-i}\nu_{n-j}}{\gamma} + \frac{\nu_i \nu_j}{\gamma} \\ &= b_{n-j,n-i} - \frac{1}{\gamma}(\nu_i \nu_j - \nu_{n-i}\nu_{n-j}). \end{aligned}$$

It needs $\frac{7}{4}n^2$ flops. See Algorithm Trench (1964) in Matrix Computations, pp.132 for details.

2.4.2 Banded Systems

Definition 2.4.2 *Let A be a $n \times n$ matrix. A is called a (p,q)-banded matrix, if $a_{ij} = 0$ whenever $i - j > p$ or $j - i > q$. A has the form*

$$A = \begin{bmatrix} \times & \cdots & \times & & O \\ \vdots & \ddots & & \ddots & \\ \times & & \ddots & & \times \\ & \ddots & & \ddots & \vdots \\ O & & \times & \cdots & \times \end{bmatrix} \begin{matrix} \top \\ q \\ \perp \end{matrix},$$

$$\underbrace{\qquad\qquad}_{p}$$

where p and q are the lower and upper band widthes, respectively.

Example 2.4.1 $(1,1)$*: tridiagonal matrix;* $(1,n-1)$*: upper Hessenberg matrix;* $(n-1,1)$*: lower Hessenberg matrix.*

Theorem 2.4.1 *Let A be a (p,q)-banded matrix. Suppose A has a LR-factorization $(A = LR)$. Then $L = (p,0)$ and $R = (0,q)$-banded matrix, respectively.* ■

Algorithm 2.4.2 *See Algorithm 4.3.1 in Matrix Computations, pp.150.*

Theorem 2.4.2 *Let A be a (p,q)-banded nonsingular matrix. If Gaussian Elimination with partial pivoting is used to compute Gaussian transformations $L_j = I - l_j e_j^T$, for $j = 1,\ldots, n-1$, and permutations $P_1, \ldots, P_{n-1}$ such that*

$$L_{n-1}P_{n-1}\cdots L_1 P_1 A = R$$

is upper triangular, then R is a $(0, p+q)$-banded matrix and $l_{ij} = 0$ whenever $i \le j$ or $i > j + p$. (Since the j-th column of L is a permutation of the Gaussian vector l_j, it follows that L has at most $p+1$ nonzero elements per column.)

2.4.3 Symmetric Indefinite Systems

Consider the linear system $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is symmetric but indefinite. There are a method using $n^3/6$ flops due to Aasen (1971) that computes the factorization $PAP^T = LTL^T$, where $L = [l_{ij}]$ is unit lower triangular, P is a permutation chosen such that $\mid l_{ij} \mid \leq 1$, and T is tridiagonal.

Rather than the above factorization $PAP^T = LTL^T$ we have the calculation of

$$PAP^T = LDL^T,$$

where D is block diagonal with 1 by 1 and 2 by 2 blocks on diagonal, $L = [l_{ij}]$ is unit lower triangular, and P is a permutation chosen such that $\mid l_{ij} \mid \leq 1$.

Bunch and Parlett (1971) has proposed a pivot strategy to do this, $n^3/6$ flops are required. Unfortunately the overall process requires $n^3/12 \sim n^3/6$ comparisons. A better method described by Bunch and Kaufmann (1977) requires $n^3/6$ flops and $O(n^2)$ comparisons.

A detailed discussion of this subsection see p.159-168 in Matrix Computations.

Chapter 3

Orthogonalization and least squares methods

3.1 QR-factorization (QR-decomposition)

3.1.1 Householder transformation

Definition 3.1.1 *A complex $m \times n$-matrix $R = [r_{ij}]$ is called an upper (lower) triangular matrix, if $r_{ij} = 0$ for $i > j$ $(i < j)$.*

Example 3.1.1

$$(1)\, m = n \;:\; R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}, \quad (2)\, m < n \;:\; R = \begin{bmatrix} r_{11} & \cdots & \cdots & \cdots & r_{1n} \\ & \ddots & & & \vdots \\ 0 & & r_{mm} & \cdots & r_{mn} \end{bmatrix},$$

$$(3)\; m > n \;:\; R = \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline & 0 & \end{array}\right].$$

Definition 3.1.2 *Given $A \in \mathbb{C}^{m\times n}, Q \in \mathbb{C}^{m\times m}$ unitary and $R \in \mathbb{C}^{m\times n}$ upper triangular as in Examples such that $A = QR$. Then the product is called a QR-factorization of A.*

Basic problem:

Given $b \neq 0, b \in \mathbb{C}^n$. Find a vector $w \in \mathbb{C}^n$ with $w^*w = 1$ and $c \in \mathbb{C}$ such that

$$(I - 2ww^*)b = ce_1. \tag{3.1.1}$$

Solution (Householder transformation):

(1) $b = 0$: w arbitrary (in general $w = 0$) and $c = 0$.

(2) $b \neq 0$:

$$c = \begin{cases} -\frac{b_1}{|b_1|}\|b\|_2, & \text{if } b_1 \neq 0, \\ \|b\|_2, & \text{if } b_1 = 0, \end{cases} \tag{3.1.2}$$

$$\begin{cases} w = \frac{1}{2k}(b_1 - c, b_2, \ldots, b_n)^T := \frac{1}{2k}u \\ \text{with} 2k = \sqrt{2\|b\|_2(\|b\|_2 + |b_1|)} \end{cases} \tag{3.1.3}$$

Theorem 3.1.1 *Any complex $m \times n$ matrix A can be factorized by the product $A = QR$, where Q is $m \times m$-unitary. R is $m \times n$ upper triangular.*

Proof: Let $A^{(0)} = A = [a_1^{(0)}|a_2^{(0)}|\cdots|a_n^{(0)}]$. Find $Q_1 = (I - 2w_1w_1^*)$ such that $Q_1a_1^{(0)} = ce_1$. Then

$$A^{(1)} = Q_1A^{(0)} = [Q_1a_1^{(0)}, Q_1a_2^{(0)}, \cdots, Q_1a_n^{(0)}] = \left[\begin{array}{c|ccc} c_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & a_2^{(1)} & \cdots & a_n^{(1)} \\ 0 & & & \end{array}\right]. \tag{3.1.4}$$

Find $Q_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & I - w_2w_2^* \end{array}\right]$ such that $(I - 2w_2w_2^*)a_2^{(1)} = c_2e_1$. Then

$$A^{(2)} = Q_2A^{(1)} = \left[\begin{array}{cc|ccc} c_1 & * & * & \cdots & * \\ 0 & c_2 & * & \cdots & * \\ \hline 0 & 0 & & & \\ \vdots & \vdots & a_3^{(2)} & \cdots & a_n^{(2)} \\ 0 & 0 & & & \end{array}\right].$$

We continue this process. Then after $l = \min(m, n)$ steps $A^{(l)}$ is an upper triangular matrix satisfying

$$A^{(l-1)} = R = Q_{l-1}\cdots Q_1A.$$

Then $A = QR$, where $Q = Q_1^*\cdots Q_{l-1}^*$. ■

Remark 3.1.1 *We usually call the method in Theorem 3.1.1 as Householder method. (Algorithm **??**).*

Theorem 3.1.2 *Let A be a nonsingular $n \times n$ matrix. Then the QR- factorization is essentially unique. That is, if $A = Q_1R_1 = Q_2R_2$, then there is a unitary diagonal matrix $D = diag(d_i)$ with $|d_i| = 1$ such that $Q_1 = Q_2D$ and $DR_1 = R_2$.*

Proof: Let $A = Q_1R_1 = Q_2R_2$. Then $Q_2^*Q_1 = R_2R_1^{-1} = D$ must be a diagonal unitary matrix. ■

Remark 3.1.2 *The QR-factorization is unique, if it is required that the diagonal elements of R are positive.*

Corollary 3.1.1 *A is an arbitrary $m \times n$-matrix. The following factorizations exist:*

(i) $A = LQ$, where Q is $n \times n$ unitary and L is $m \times n$ lower triangular.

(ii) $A = QL$, where Q is $m \times m$ unitary and L is $m \times n$ lower triangular.

(iii) $A = RQ$, where Q is $n \times n$ unitary and R is $m \times n$ upper triangular.

Proof: (i) A^* has a QR-factorization. Then

$$A^* = QR \Rightarrow A = R^*Q^* \Rightarrow (i).$$

(ii) Let $P_m = \begin{bmatrix} O & 1 \\ 1 & O \end{bmatrix}$. Then by Theorem 3.1.1 we have $P_mAP_n = QR$. This implies

$$A = (P_mQP_m)(P_mRP_n) \equiv \tilde{Q}L \Rightarrow (ii).$$

(iii) A^* has a QL-factorization (from (ii)), i.e., $A^* = QL$. This implies

$$A = L^*Q^* \Rightarrow (iii).$$

■

Cost of Householder method

Consider that the multiplications in (3.1.4) can be computed in the form

$$\begin{aligned}(I - 2w_1w_1^*)A &= (I - \frac{u_1}{\|b\|_2^2 + |b_1|\|b\|_2}u_1^*)A = (I - vu_1^*)A \\ &= A - vu_1^*A := A - vw^*.\end{aligned}$$

So the first step for a $m \times n$-matrix A requires;

c_1: m multiplications, 1 root;
$4k^2$: 1 multiplication;
v: m divisions (= multiplications);
w: mn multiplications;
$A^{(1)} = A - vw^*$: $m(n-1)$ multiplications.

Similarly, for the j-th step m and n are replaced by $m-j+1$ and $n-j+1$, respectively. Let $l = \min(m, n)$. Then the number of multiplications is

$$\sum_{j=1}^{l-1}[2(m-j+1)(n-j+1) + (m-j+2)] \tag{3.1.5}$$
$$= l(l-1)[\frac{2l-1}{3} - (m+n) - 5/2] + (l-1)(2mn + 3m + 2n + 4)$$
$$(= mn^2 - 1/3n^3, \text{ if } m \geq n).$$

Especially, for $m = n$, it needs

$$\sum_{j=1}^{n-1}[2(n-j+1)^2 + m - j + 2] = 2/3n^3 + 3/2n^2 + 11/6n - 4 \tag{3.1.6}$$

flops and $(l + n - 2)$ roots. To compute $Q = Q_1^* \cdots Q_{l-1}^*$, it requires

$$2[m^2n - mn^2 + n^3/3] \text{ multiplications } (m \geq n). \tag{3.1.7}$$

Remark 3.1.3 *Let $A = QR$ be a QR-factorization A. Then we have*

$$A^*A = R^*Q^*QR = R^*R.$$

If A has full column rank and we require that the diagonal elements of R are positive, then we obtain the Cholesky factorization of A^*A.

3.1.2 Gram-Schmidt method

Remark 3.1.4 *Theorem 3.1.1 (or Algorithm **??**) can be used to solved orthonormal basis (OB) problem.*

(OB) *: Given linearly independent vectors $a_1, \cdots, a_n \in \mathbb{R}^{n\times 1}$. Find an orthonormal basis for span$\{a_1, \cdots, a_n\}$.*

If $A = [a_1, \cdots, a_n] = QR$ with $Q = [q_1, \cdots, q_n]$, and $R = [r_{ij}]$, then

$$a_k = \sum_{i=1}^{k} r_{ik} q_i. \tag{3.1.8}$$

By assumption rank$(A) = n$ and (3.1.8) it implies $r_{kk} \neq 0$. So, we have

$$q_k = \frac{1}{r_{kk}}(a_k - \sum_{i=1}^{k-1} r_{ik} q_i). \tag{3.1.9}$$

The vector q_k can be thought as a unit vector in the direction of $z_k = a_k - \sum_{i=1}^{k-1} s_{ik} q_i$. To ensure that $z_k \perp q_1, \cdots, q_{k-1}$ we choose $s_{ik} = q_i^T a_k$, for $i = 1, \cdots, k-1$. This leads to the Classical Gram-Schmidt (CGS) Algorithm for solving (OB) problem.

Algorithm 3.1.1 (Classical Gram-Schmidt (CGS) Algorithm) *Given $A \in \mathbb{R}^{m\times n}$ with rank$(A) = n$. We compute $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns and $R \in \mathbb{R}^{n\times n}$.*

For $i = 1, \cdots, n$,
  $q_i = a_i$;
  For $j = 1, \cdots, i-1$
    $r_{ji} = q_j^T a_i$,
    $q_i = q_i - r_{ji} q_j$,
  end for
  $r_{ii} = \|q_i\|_2$,
  $q_i = q_i / r_{ii}$,
end for

Disadvantage : The CGS method has very poor numerical properties, if some columns of A are nearly linearly independent.

Advantage : The method requires mn^2 multiplications ($m \geq n$).

Remark 3.1.5 *Modified Gram-Schmidt (MGS):*

Write $A = \sum_{i=1}^{n} q_i r_i^T$. Define $A^{(k)}$ by

$$[0, A^{(k)}] = A - \sum_{i=1}^{k-1} q_i r_i^T = \sum_{i=k}^{n} q_i r_i^T \tag{3.1.10}$$

It follows that if $A^{(k)} = [z, B], z \in \mathbb{R}^m$, $B \in \mathbf{R}^{m\times(n-k)}$ then $r_{kk} = \|z\|_2$ and $q_k = z/r_{kk}$ by (3.1.9). Compute

$$[r_{k,k+1}, \cdots, r_{kn}] = q_k^T B.$$

Next step: $A^{(k+1)} = B - q_k[r_{k,k+1}, \cdots, r_{kn}]$.

Algorithm 3.1.2 (MGS) *Given $A \in \mathbb{R}^{m\times n}$ with $rank(A) = n$. We compute $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns and $R \in \mathbb{R}^{n\times n}$ is upper triangular.*

For $i = 1, \cdots, n$,
  $q_i = a_i$;
  For $j = 1, \cdots, i-1$
    $r_{ji} = q_j^T q_i$,
    $q_i = q_i - r_{ji} q_j$,
  end for
  $r_{ii} = \|q_i\|_2$,
  $q_i = q_i / r_{ii}$,
end for

The MGS requires mn^2 multiplications.

Remark 3.1.6 *MGS computes the QR factorization at the kth step, the kth column of Q and the kth row of R are computed. CGS at the kth step, the kth column of Q and the kth column of R are computed.*

Advantage for OB problem ($m \geq n$): (i) Householder method requires $mn^2 - n^3/3$ flops to get factorization. $A = QR$ and $mn^2 - n^3/3$ flops to get the first n columns of Q. But MGS requires only mn^2 flops. Thus for the problem of finding an orthonormal basis of range(A), MGS is about twice as efficient as Householder orthogonalization. (ii) MGS is numerically stable.

3.1.3 Givens method

Basic problem: Given $(a, b)^T \in \mathbb{R}^2$, find c, $s \in \mathbb{R}$ with $c^2 + s^2 = 1$ such that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} k \\ 0 \end{bmatrix},$$

where $c = cos\alpha$ and $s = sin\alpha$.
Solution:

$$\begin{cases} c = 1,\ s = 0,\ k = a;\ \text{if } b = 0, \\ c = \frac{a}{\sqrt{a^2+b^2}},\ s = \frac{b}{\sqrt{a^2+b^2}},\ k = \sqrt{a^2+b^2};\ \text{if } b \neq 0. \end{cases} \tag{3.1.11}$$

Let

$$G(i, j, \alpha) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & \vdots & & \vdots & \\ & \cdots & \cos\alpha & \cdots & \sin\alpha & \cdots & \\ & & \vdots & & \vdots & & \\ & \cdots & -\sin\alpha & \cdots & \cos\alpha & \cdots & \\ & & \vdots & & \vdots & \ddots & \\ & & & & & & 1 \end{bmatrix}.$$

Then $G(i, j, \alpha)$ is called a Givens rotation in the (i, j)-coordinate plane. In the matrix $\tilde{A} = G(i, j, \alpha)A$, the rows with index $\neq i, j$ are the same as in A and

$$\begin{aligned} \tilde{a}_{ik} &= \cos(\alpha)a_{ik} + \sin(\alpha)a_{jk},\ \text{for } k = 1, \ldots, n, \\ \tilde{a}_{jk} &= -\sin(\alpha)a_{ik} + \cos(\alpha)a_{jk},\ \text{for } k = 1, \ldots, n. \end{aligned}$$

Algorithm 3.1.3 (Givens orthogonalization) *Given $A \in \mathbb{R}^{m\times n}$. The folllowing Algorithm overwrites A with $Q^TA = R$, where Q is orthonormal and R is upper triangular. For $q = 2, \cdots, m$,*

for $p = 1, 2, \cdots, \min\{q-1, n\}$,

Find $c = \cos\alpha$ and $s = \sin\alpha$ as in (3.1.11) such that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a_{pp} \\ a_{qp} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}.$$

$A := G(p, q, \alpha)A$.

This algorithm requires $2n^2(m - n/3)$ flops.

Fast Givens method (See Matrix Computations, pp.205-209):

A modification of Givens method bases on the fast Givens rotations and requires about $n^2(m - n/3)$ flops.

3.2 Overdetermined linear Systems - Least Squares Methods

Given $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ and $m > n$. Consider the least squares(LS) problem:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2. \tag{3.2.1}$$

Let X be the set of minimizers defined by $X = \{x \in \mathbb{R}^n \mid \|Ax - b\|_2 = \min!\}$. It is easy to see the following properties:

- $x \in X \Longleftrightarrow A^T(b - Ax) = 0.$ (3.2.2)
- X is convex. (3.2.3)
- X has a unique element x_{LS} having minimal 2-norm. (3.2.4)
- $X = \{x_{LS}\} \Longleftrightarrow rank(A) = n.$ (3.2.5)

For $x \in \mathbb{R}^n$, we refer to $r = b - Ax$ as its residual. $A^T(b - Ax) = 0$ is refered to as the normal equation. The minimum sum is defined by $\rho_{LS}^2 = \|Ax_{LS} - b\|_2^2$. If we let $\varphi(x) = \frac{1}{2}\|Ax - b\|_2^2$, then $\nabla\varphi(x) = A^T(Ax - b)$.

Theorem 3.2.1 *Let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, with r =rank(A), $U = [u_1, \ldots, u_m]$ and $V = [v_1, \cdots, v_n]$ be the SVD of $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$. If $b \in \mathbb{R}^m$, then*

$$x_{LS} = \sum_{i=1}^{r} (u_i^T b/\sigma_i) v_i \tag{3.2.6}$$

and

$$\rho_{LS}^2 = \sum_{i=r+1}^{m} (u_i^T b)^2 \tag{3.2.7}$$

Proof: For any $x \in \mathbb{R}^n$ we have

$$\|Ax - b\|_2^2 = \|U^T AV(V^T x) - U^T b\|_2^2 = \sum_{i=1}^{r}(\sigma_i \alpha_i - u_i^T b)^2 + \sum_{i=r+1}^{m}(u_i^T b)^2,$$

where $\alpha = V^T x$. Clearly, if x solves the LS-problem, then $\alpha_i = (u_i^T b/\sigma_i)$, for $i = 1, \ldots, r$. If we set $\alpha_{r+1} = \cdots = \alpha_n = 0$, then $x = x_{LS}$. ■

Remark 3.2.1 If we define A^+ by $A^+ = V\Sigma^+ U^T$, where $\Sigma^+ = \text{diag}(\sigma_1^{-1}, .., \sigma_r^{-1}, 0, .., 0)$ $\in \mathbb{R}^{n\times m}$ then $\boxed{x_{\text{LS}} = A^+ b}$ and $\rho_{\text{LS}} = \|(I - AA^+)b\|_2$. A^+ is refered to as the pseudo-inverse of A. A^+ is defined to be the unique matrix $X \in \mathbb{R}^{n\times m}$ that satisfies Moore-Penrose conditions :

$$\begin{array}{ll} (i)AXA = A, & (iii)\ (AX)^T = AX, \\ (ii)XAX = X, & (iv)\ (XA)^T = XA. \end{array} \tag{3.2.8}$$

Existence of X is easy to check by taking $X = A^+$. Now, we show the uniqueness of X. Suppose X and Y satisfying the conditions (i)–(iv). Then

$$\begin{aligned} X &= XAX = X(AYA)X = X(AYA)Y(AYA)X \\ &= (XA)(YA)Y(AY)(AX) = (XA)^T(YA)^T Y(AY)^T(AX)^T \\ &= (AXA)^T Y^T YY^T (AXA)^T = A^T Y^T YY^T A^T \\ &= Y(AYA)Y = YAY = Y. \end{aligned}$$

If $\text{rank}(A) = n$ $(m \geq n)$, then $A^+ = (A^T A)^{-1} A^T$. If $\text{rank}(A) = m$ $(m \leq n)$, then $A^+ = A^T(AA^T)^{-1}$. If $m = n = \text{rank}(A)$, then $A^+ = A^{-1}$.

• For the case rank(A)=n:

Algorithm 3.2.1 (Normal equations) *Given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ with rank(A) = n and $b \in \mathbb{R}^m$. This Algorithm computes the solution to the LS-problem:* $\min\{\|Ax - b\|_2; x \in \mathbb{R}^n\}$.
Compute $d := A^T b$, and form $C := A^T A$ by computing the Cholesky factorization $C = R^T R$ (see Remark 6.1). Solve $R^T y = d$ and $Rx_{LS} = y$.

Algorithm 3.2.2 (Householder and Givens orthogonalizations) *Given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ with rank(A) = n and $b \in \mathbb{R}^m$. This Algorithm computes the solutins to the LS-problem:* $\min\{\|Ax - b\|_2; x \in \mathbb{R}^n\}$.
Compute QR-factorization $Q^T A = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$ by using Householder and Givens methods respectively. (Here R_1 is upper triangular). Then

$$\|Ax - b\|_2^2 = \|Q^T Ax - Q^T b\|_2^2 = \|R_1 x - c\|_2^2 + \|d\|_2^2,$$

where $Q^T b = \begin{bmatrix} c \\ d \end{bmatrix}$. Thus, $x_{LS} = R_1^{-1} c$, (since rank(A) =rank(R_1) = n) and $\rho_{LS}^2 = \|d\|_2^2$.

Algorithm 3.2.3 (Modified Gram-Schmidt) *Given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ with rank$(A) = n$ and $b \in \mathbb{R}^m$. The solution of* $\min \|Ax - b\|_2$ *is given by:*
Compute $A = Q_1 R_1$, where $Q_1 \in \mathbb{R}^{m\times n}$ with $Q_1^T Q_1 = I_n$ and $R_1 \in \mathbb{R}^{n\times n}$ upper triangular. Then the normal equation $(A^T A)x = A^T b$ is transformed to the linear system $R_1 x = Q_1^T b \Rightarrow x_{LS} = R_1^{-1} Q_1^T b$.

• For the case rank$(A) < n$:
Problem:

(i) How to find a solution to the LS-problem?

(ii) How to find the unique solution having minimal 2-norm?

(iii) How to compute x_{LS} reliably with infinite conditioned A ?

Definition 3.2.1 *Let A be a $m \times n$ matrix with rank$(A) = r$ $(r \leq m, n)$. The factorization $A = BC$ with $B \in \mathbb{R}^{m\times r}$ and $C \in \mathbb{R}^{r\times n}$ is called a full rank factorization, provided that B has full column rank and C has full row rank.*

Theorem 3.2.2 *If $A = BC$ is a full rank factorization, then*

$$A^+ = C^+ B^+ = C^T (CC^T)^{-1} (B^T B)^{-1} B^T. \tag{3.2.9}$$

Proof: From assumption follows that

$$\begin{aligned} B^+ B &= (B^T B)^{-1} B^T B = I_r, \\ CC^+ &= CC^T (CC^T)^{-1} = I_r. \end{aligned}$$

We calculate (3.2.8) with

$$\begin{aligned} &A(C^+ B^+)A = BCC^+ B^+ BC = BC = A, \\ &(C^+ B^+)A(C^+ B^+) = C^+ B^+ BCC^+ B^+ = C^+ B^+, \\ &A(C^+ B^+) = BCC^+ B^+ = BB^+ \quad \text{symmetric}, \\ &(C^+ B^+)A = C^+ B^+ BC = C^+ C \quad \text{symmetric}. \end{aligned}$$

These imply that $X = C^+ B^+$ satisfies (3.2.8). It follows $A^+ = C^+ B^+$. ■

Unfortunately, if rank$(A) < n$, then the QR-factorization does not necessarily produce a full rank factorization of A. For example

$$A = [a_1, a_2, a_3] = [q_1, q_2, q_3] \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Fortunately,we have the following two methods to produce a full rank factorization of A.

3.2.1 Rank Deficiency I : QR with column pivoting

Algorithm **??** can be modified in a simple way so as to produce a full rank factorization of A.

$$A\Pi = QR,\ R = \underbrace{\begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}}_{\begin{matrix} r & n-r \end{matrix}} \begin{matrix} \}\text{r} \\ \}\text{m-r} \end{matrix}, \tag{3.2.10}$$

where $r = \text{rank}(A) < n$ $(m \geq n)$, Q is orthogonal, R_{11} is nonsingular upper triangular and Π is a permuatation. Once (3.2.10) is computed, then the LS-problem can be readily solved by

$$\|Ax - b\|_2^2 = \|(Q^T A\pi)(\pi^T x) - Q^T b\|_2^2 = \|R_{11}y - (c - R_{12}z)\|_2^2 + \|d\|_2^2,$$

where $\Pi^T x = \begin{bmatrix} y \\ z \end{bmatrix} \begin{matrix} \}\text{r} \\ \}\text{n-r} \end{matrix}$ and $Q^T b = \begin{bmatrix} c \\ d \end{bmatrix} \begin{matrix} \}\text{r} \\ \}\text{m-r} \end{matrix}$. Thus if $\|Ax - b\|_2 = \min!$, then we must have

$$x = \Pi \begin{bmatrix} R_{11}^{-1}(c - R_{12}z) \\ z \end{bmatrix}.$$

If z is set to be zero, then we obtain the basic solution

$$x_{\text{B}} = \pi \begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix}.$$

The basic solution is not the solution with minimal 2-norm, unless the submatrix R_{12} is zero. Since

$$\|x_{\text{LS}}\|_2 = \min_{z \in \mathbb{R}^{n-r}} \left\| x_{\text{B}} - \pi \begin{bmatrix} R_{11}^{-1}R_{12} \\ -I_{n-r} \end{bmatrix} z \right\|_2. \tag{3.2.11}$$

We now solve the LS-problem (3.2.11) by using Algorithms 3.2.1 to 3.2.3.

Algorithm 3.2.4 *Given $A \in \mathbb{R}^{m \times n}$, with rank$(A) = r < n$. The following algorithm computes the factorization $A\Pi = QR$ defined by (3.2.10). The element a_{ij} is overwritten by r_{ij} $(i \leq j)$. The permutation $\Pi = [e_{c_1}, \ldots, e_{c_n}]$ is determined according to choosing the maximum of column norm in the current step.*

$c_j := j$ $(j = 1, 2, \ldots, n)$,
$r_j := \sum_{i=1}^{m} a_{ij}^2$ $(j = 1, \ldots, n)$,
For $k = 1, \ldots, n$,
  Detemine p with $(k \leq p \leq n)$ so that $r_p = \max_{k \leq j \leq n} r_j$.
  If $r_p = 0$ then stop; else
  Interchange c_k and c_p, r_k and r_p, and a_{ik} and a_{ip}, for $i = 1, \ldots, m$.
  Determine a Householder $\hat{Q}_k$ such that

$$\hat{Q}_k \begin{bmatrix} a_{kk} \\ \vdots \\ \vdots \\ a_{mk} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$A := diag(I_{k-1}, \hat{Q}_k)A$; $r_j := r_j - a_{kj}^2 (j = k + 1, \ldots, n)$.

This algorithm requires $2mnr - r^2(m+n) + 2r^3/3$ flops.

Algorithm 3.2.4 produces the full rank factorization (3.2.10) of A. We have the following important relations:

$$\begin{cases} |r_{11}| \geq |r_{22}| \geq \ldots \geq |r_{rr}|, \ r_{jj} = 0, \ j = r+1, \ldots, n, \\ |r_{ii}| \geq |r_{ik}|, \ i = 1, \ldots, r, \ k = i+1, \ldots, n. \end{cases} \tag{3.2.12}$$

Here, $r = \text{rank}(A) < n$, and $R = (r_{jj})$. In the following we show another application of the full rank factorization for solving the LS-problem.

Algorithm 3.2.5 (Compute $x_{\mathbf{LS}} = A^+ b$ directly)

(i) Compute (3.2.10): $A\Pi = QR \equiv (\underbrace{Q^{(1)}}_{r} \mid Q^{(2)}) \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \begin{matrix} \}r \\ \}m\text{-}r \end{matrix}, \Rightarrow \ A\Pi = Q^{(1)} R_1.$

(ii) $(A\Pi)^+ = R_1^+ Q^{(1)^+} = R_1^+ Q^{(1)^T}.$

(iii) Compute R_1^+*:*

Either: $R_1^+ = R_1^T (R_1 R_1^T)^{-1}$ *(since* R_1 *has full row rank)*
$\Rightarrow (A\Pi)^+ = R_1^T (R_1 R_1^T)^{-1} Q^{(1)^T}.$

Or: Find $\widehat{\dot{Q}}$ *using Householder transformation (Algorithm* **??***) such that* $\widehat{\dot{Q}} R_1^T = \begin{bmatrix} T \\ 0 \end{bmatrix}$*, where* $T \in \mathbb{R}^{r \times r}$ *is upper triangular.*

Let $\widehat{Q}^T := (\hat{Q}^{(1)}, \hat{Q}^{(2)}) \ \Rightarrow R_1^T = \hat{Q}^{(1)} T + \hat{Q}^{(2)} 0 = \hat{Q}^{(1)} T.$
$R_1 = T^T \hat{Q}^{(1)^T} \Rightarrow R_1^+ = (\hat{Q}^{(1)T})^+ (T^T)^+ = \hat{Q}^{(1)} (T^T)^{-1}.$
$\Rightarrow (A\Pi)^+ = \hat{Q}^{(1)} (T^T)^{-1} Q^{(1)^T}.$

(iv) Since $\min \|Ax - b\|_2 = \min \|A\Pi(\Pi^T x) - b\|_2 \Rightarrow (\Pi^T x)_{LS} = (A\Pi)^+ b$
$\Rightarrow \boxed{x_{LS} = \Pi (A\Pi)^+ b}.$

Remark 3.2.2 Unfortunately, QR with column pivoting is not entirely reliable as a method for detecting near rank deficiency. For example:

$$T_n(c) = \text{diag}(1, s, \cdots, s^{n-1}) \begin{bmatrix} 1 & -c & -c & \cdots & -c \\ & 1 & -c & \cdots & -c \\ & & \ddots & & \vdots \\ 0 & & & & 1 \end{bmatrix} \quad c^2 + s^2 = 1, c, s > 0.$$

If $n = 100$, $c = 0.2$, then σ_n=0.3679e$-$8. But this matrix is unaltered by Algorithm 3.2.4. However,the "degree of unreliability" is somewhat like that for Gaussian elimination with partial pivoting, a method that works very well in practice.

3.2.2 Rank Deficiency II : The Singular Value Decomposition

Algorithm 3.2.6 (Householder Bidiagonalization) *Given $A \in \mathbb{R}^{m\times n}\,(m \geq n)$. The following algorithm overwrite A with $U_B^T A V_B = B$, where B is upper bidiagonal and U_B and V_B are orthogonal.*
For $k = 1, \cdots, n$,

Determine a Householder matrix $\tilde{U}_k$ of order $n-k+1$ such that

$$\hat{\dot{U}}_k \begin{bmatrix} a_{kk} \\ \vdots \\ \vdots \\ a_{mk} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$A := diag(I_{k-1}, \hat{\dot{U}}_k)A$,

If $k \leq 2$, then determine a Householder matrix $\hat{\dot{V}}_k$ of order $n-k+1$ such that

$$[a_{k,k+1}, \cdots, a_{kn}]\hat{\dot{V}}_k = (*, 0, \cdots, 0),$$

$A := Adiag(I_k, \hat{\dot{V}}_k)$.

This algorithm requires $2mn^2 - 2/3n^3$ flops.

Algorithm 3.2.7 (R-Bidiagonalization) *when $m \gg n$ we can use the following faster method of bidiagonalization.*

(1) Compute an orthogonal $Q_1 \in \mathbb{R}^{m\times m}$ such that $Q_1^T A = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$, where $R_1 \in \mathbb{R}^{n\times n}$ is upper triangular.

(2) Applying Algorithm 3.2.6 to R_1, we get $Q_2^T R_1 V_B = B_1$, where Q_2, $V_B \in \mathbb{R}^{n\times n}$ orthogonal and $B_1 \in \mathbb{R}^{n\times n}$ upper bidiagonal.

(3) Define $U_B = Q_1 diag(Q_2, I_{m-n})$. Then $U_B^T A V_B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \equiv B$ bidiagonal.

This algorithm require $mn^2 + n^3$. It involves fewer compuations comparing with Algorithm 7.6 ($2mn^2 - 2/3n^3$) whenever $m \geq 5/3n$.

Once the bidiagonalization of A has been achieved,the next step in the Golub-Reinsch SVD algorithm is to zero out the super diagonal elements in B. Unfortunately, we must defer our discussion of this iteration until Chapter 5 since it requires an understanding of the symmetric QR algorithm for eigenvalues. That is, it computes orthogonal matrices U_Σ and V_Σ such that

$$U_\Sigma^T B V_\Sigma = \Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_n).$$

By defining $U = U_B U_\Sigma$ and $V = V_B V_\Sigma$, we see that $U^T A V = \Sigma$ is the SVD of A.

Algorithms			Flop Counts
rank(A)=n	Algorithm 3.2.1	Normal equations	$mn^2/2 + n^3/6$
	Algorithm 3.2.2	Householder orthogonalization	$mn^2 - n^3/3$
	Algorithm 3.2.3	Modified Gram-Schmidt	mn^2
	Algorithm 3.1.3	Givens orthogonalization	$2mn^2 - 2/3n^3$
	Algorithm 3.2.6	Householder Bidiagonalization	$2mn^2 - 2/3n^3$
	Algorithm 3.2.7	R-Bidiagonalization	$mn^2 + n^3$
rank(A) $< n$	LINPACK	Golub-Reinsch SVD	$2mn^2 + 4n^3$
	Algorithm 3.2.5	QR-with column pivoting	$2mnr - mr^2 + 1/3r^3$
	Alg. 3.2.7+SVD	Chan SVD	$mn^2 + 11/2n^3$

Table 3.1: Solving the LS problem ($m \geq n$)

Remark 3.2.3 *If the* LINPACK *SVD Algorithm is applied with eps*=10^{-17} *to*

$$T_{100}(0.2) = diag(1, s, \cdots, s^{n-1}) \begin{bmatrix} 1 & -c & -c & \cdots & -c \\ & 1 & -c & \cdots & -c \\ & & \ddots & & \vdots \\ 0 & & & & 1 \end{bmatrix},$$

then $\hat{\sigma}_n = 0.367805646308792467 \times 10^{-8}$.

Remark 3.2.4 *As we mentioned before, when solving the LS problem via the SVD, only* Σ *and* V *have to be computed (see (3.2.6)). Table 3.1 compares the efficiency of this approach with the other algorithms that we have presented.*

3.2.3 The Sensitivity of the Least Squares Problem

Corollary 3.2.1 (of Theorem 1.2.3) *Let* $U = [u_1, \cdots, u_m]$, $V = [v_1, \cdots, v_n]$ *and* $U^*AV = \Sigma = diag(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$. *If* $k < r = rank(A)$ *and* $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, *Then*

$$\min_{rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Proof: Since $U^T A_k V = \text{diag}(\sigma_1, \cdots, \sigma_k, 0, \cdots, 0)$, it follows $\text{rank}(A_k) = k$ and that

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\text{diag}(0, \cdots, 0, \sigma_{k+1}, \cdots, \sigma_r)\|_2 = \sigma_{k+1}.$$

Suppose $B \in \mathbb{R}^{m\times n}$ and $\text{rank}(B) = k$, i.e., there are orthogonal vectors $x_1, \cdots, x_{n-k}$ such that $\mathcal{N}(B) = span\{x_1, \cdots, x_{n-k}\}$. This implies

$$span\{x_1, \cdots, x_{n-k}\} \bigcap span\{v_1, \cdots, v_{k+1}\} \neq \{0\}.$$

Let z be a unit vector in the intersection set. Then $Bz = 0$ and $Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$. Thus,

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2.$$

■

3.2.4 Condition number of a Rectangular Matrix

Let $A \in \mathbb{R}^{m \times n},\ \mathrm{rank}(A) = n,\ \kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$.
(i) The method of normal equation:

$$\min_{x \in R^n} \|Ax - b\|_2 \Leftrightarrow A^T A x = A^T b.$$

(a) $C = A^T A,\ d = A^T b$.

(b) Compute the Cholesky factorization $C = GG^T$.

(c) Solve $Gy = d$ and $G^T x_{\mathrm{LS}} = y$. Then

$$\frac{\|\tilde{x}_{LS} - x_{\mathrm{LS}}\|_2}{x_{\mathrm{LS}}} \approx \mathrm{eps}\kappa_2(A^T A) = \mathrm{eps}\kappa_2(A)^2.$$

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A) \left(\varepsilon \frac{\|F\|}{\|A\|} + \varepsilon \frac{\|f\|}{\|b\|} \right) + o(\varepsilon^2),$$

where $(A + F)\tilde{x} = b + f$ and $Ax = b$.

(ii) LS solution via QR factorization

$$\|Ax - b\|_2^2 = \|Q^T Ax - Q^T b\|_2^2 = \|R_1 x - c\|_2^2 + \|d\|_2^2,$$

$$x_{\mathrm{LS}} = R_1^{-1} c,\ \rho_{\mathrm{LS}} = \|d\|_2.$$

Numerically, trouble can be expected wherever $\kappa_2(A) = \kappa_2(R) \approx 1/eps$. But this is in contrast to normal equation, Cholesky factorization becomes problematical once $\kappa_2(A)$ is in the neighborhood of $1/\sqrt{eps}$.

Remark 3.2.5

$$\begin{aligned} \|A\|_2 \|(A^T A)^{-1} A^T\|_2 &= \kappa_2(A), \\ \|A\|_2^2 \|(A^T A)^{-1}\|_2 &= \kappa_2(A)^2. \end{aligned}$$

Theorem 3.2.3 *Let $A \in \mathbb{R}^{m \times n}, (m \geq n),\ b \neq 0$. Suppose that $x,\ r,\ \tilde{x},\ \tilde{r}$ satisfy*

$$\begin{aligned} &\|Ax - b\| = \min!,\ r = b - Ax,\ \rho_{LS} = \|r\|_2, \\ &\|(A + \delta A)\tilde{x} - (b + \delta b)\|_2 = \min!, \\ &\tilde{r} = (b + \delta b) - (A + \delta A)\tilde{x}. \end{aligned}$$

If

$$\varepsilon = \max\{\frac{\|\delta A\|_2}{\|A\|_2}, \frac{\|\delta b\|_2}{\|b\|_2}\} < \frac{\sigma_n(A)}{\sigma_1(A)}$$

and

$$\sin\theta = \frac{\rho_{LS}}{\|b\|_2} \neq 1,$$

then

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leq \varepsilon\{\frac{2\kappa_2(A)}{\cos\theta} + \tan\theta\kappa_2(A)^2\} + O(\varepsilon^2)$$

and

$$\frac{\|\tilde{r} - r\|_2}{\|b\|_2} \leq \varepsilon(1 + 2\kappa_2(A))\min(1, m-n) + O(\varepsilon^2).$$

Proof: Let $E = \delta A/\varepsilon$ and $f = \delta b/\varepsilon$. Since $\|\delta A\|_2 < \sigma_n(A)$, by previous Corollary follows that $\text{rank}(A + \varepsilon E) = n$ for $t \in [0, \varepsilon]$.
[$t = \varepsilon \Rightarrow A + tE = A + \delta A$. If $\text{rank}(A + \delta A) = k < n$, then $\|A - (A + \delta A)\|_2 = \|\delta A\|_2 \geq \|A - A_k\|_2 = \sigma_{k+1} \geq \delta_n$. Contradiction! So $\min\limits_{rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2$ $= \|A - \sum\limits_{i=1}^{k} \sigma_i u_i v_i^T\|_2 = \sigma_{k+1}$].
Hence we have,

$$(A + tE)^T(A + tE)x(t) = (A + tE)^T(b + tf). \tag{3.2.13}$$

Since $x(t)$ is continuously differentiable for all $t \in [0, \varepsilon]$, $x = x(0)$ and $\tilde{x} = \lambda(\varepsilon)$, it follows that

$$\tilde{x} = x + \varepsilon\dot{x}(0) + O(\varepsilon^2)$$

and

$$\frac{\|\tilde{x} - x\|}{\|x\|} = \varepsilon\frac{\|\dot{x}(0)\|_2}{\|x\|} + O(\varepsilon^2).$$

Differentiating (3.2.13) and setting $t = 0$ then we have

$$E^T Ax + A^T Ex + A^T A\dot{x}(0) = A^T f + E^T b.$$

Thus,

$$\dot{x}(0) = (A^T A)^{-1}A^T(f - Ex) + (A^T A)^{-1}E^T r.$$

From $\|f\|_2 \leq \|b\|_2$ and $\|E\|_2 \leq \|A\|_2$ follows

$$\begin{aligned}\frac{\|\tilde{x} - x\|_2}{\|x\|_2} &\leq \varepsilon\{\|A\|_2\|(A^T A)^{-1}A^T\|_2(\frac{\|b\|_2}{\|A\|_2\|x\|_2} + 1) \\ &+ \frac{\rho_{LS}}{\|A\|_2\|x\|_2}\|A\|_2^2\|(A^T A)^{-1}\|_2\} + O(\varepsilon^2).\end{aligned}$$

Since $A^T(Ax_{LS} - b) = 0$, $Ax_{LS} \perp Ax_{LS} - b$ and then

$$\|b - Ax\|_2^2 + \|Ax\|_2^2 = \|b\|_2^2$$

and

$$\|A\|_2^2\|x\|_2^2 \geq \|b\|_2^2 - \rho_{LS}^2.$$

Thus,

$$\frac{\|\tilde{x}-x\|_2}{\|x\|_2} \leq eps\{\kappa_2(A)(\frac{1}{\cos\theta}+1)+\kappa_2(A)^2\frac{\sin\theta}{\cos\theta}\}+O(\varepsilon^2).$$

Furthermore, by $\frac{\sin\theta}{\cos\theta}=\frac{\rho_{LS}}{\sqrt{\|b\|_2^2-\rho_{LS}^2}}$, we have

$$\frac{\|\tilde{x}-x\|_2}{\|x\|_2} \approx eps(\kappa_2(A)+\kappa_2(A)^2\rho_{LS}). \quad (\theta: \text{ small })$$

■

Remark 3.2.6 Normal equation: $eps\ \kappa_2(A)^2$.
QR-approach: $eps(\kappa_2(A)+\rho_{LS}\kappa_2(A)^2)$.

(i) If ρ_{LS} is small and $\kappa_2(A)$ is large, then QR is better than the normal equation.

(ii) The normal equation approach involves about half of the arithmetic when $m \gg n$ and does not requires as much storage.

(iii) The QR approach is applicable to a wider class of matrices because the Cholesky to A^TA break down "before" the back substitution process on $Q^TA=R$.

3.2.5 Iterative Improvement

$$\begin{bmatrix} I_m & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} r \\ x \end{bmatrix}=\begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \|b-Ax\|_2=min!$$

$r+Ax=b, \quad A^Tr=0 \Rightarrow A^TAx=A^Tb$. Thus,

$$\begin{bmatrix} f^{(k)} \\ g^{(k)} \end{bmatrix}=\begin{bmatrix} b \\ 0 \end{bmatrix}-\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} r^{(k)} \\ x^{(k)} \end{bmatrix} \text{ and } \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} p^{(k)} \\ z^{(k)} \end{bmatrix}=\begin{bmatrix} f^{(k)} \\ g^{(k)} \end{bmatrix}.$$

This implies,

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

If $A=QR=Q\begin{bmatrix} R_1 \\ 0 \end{bmatrix}$, then $\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} p \\ z \end{bmatrix}=\begin{bmatrix} f \\ g \end{bmatrix}$ implies that

$$\left[\begin{array}{cc|c} I_n & 0 & R_1 \\ 0 & I_{m-n} & 0 \\ \hline R_1^T & 0 & 0 \end{array}\right]\begin{bmatrix} h \\ f_2 \\ z \end{bmatrix}=\begin{bmatrix} f_1 \\ f_2 \\ g \end{bmatrix},$$

where $Q^Tf=\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}Q^Tp=\begin{bmatrix} h \\ f_2 \end{bmatrix}$. Thus, $R_1^Th=g \Rightarrow h=R_1^{-T}g$. Then

$$z=R_1^{-1}(f_1-h), \quad P=Q\begin{bmatrix} h \\ f_2 \end{bmatrix}.$$

■

Chapter 4

Iterative Methods for Solving Large Linear Systems

4.1 General procedures for the construction of iterative methods

Given a linear system of nonsingular A

$$Ax = b. \tag{4.1.1}$$

Let

$$A = F - G \tag{4.1.2}$$

with F nonsingular. Then (4.1.1) is equivalent to $Fx = Gx + b$; or letting $T = F^{-1}G$ and $f = F^{-1}b$ we have

$$x = Tx + f. \tag{4.1.3}$$

Set

$$x^{(k+1)} = Tx^{(k)} + f, \tag{4.1.4}$$

where $x^{(0)}$ is given. Then the solution x of (4.1.1) is determined by iteration.

Example 4.1.1 We consider the standard decomposition of A

$$A = D - L - R, \tag{4.1.5}$$

where $A = [a_{ij}]_{i,j=1}^{n}$,

$$
\begin{aligned}
D &= \operatorname{diag}(a_{11}, a_{22}, \cdots, a_{nn}), \\
-L &= \begin{bmatrix} 0 & & & & 0 \\ a_{2,1} & 0 & & & \\ a_{3,1} & a_{3,2} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n-1} & 0 \end{bmatrix}, \\
-R &= \begin{bmatrix} 0 & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ & 0 & a_{2,3} & \cdots & a_{2,n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & a_{n-1,n} \\ 0 & & & \ddots & 0 \end{bmatrix}.
\end{aligned}
$$

For $a_{i,i} \neq 0, i = 1, \cdots, n$, D is nonsingular. If we choose $F = D$ and $G = L + R$ in (8.2), we then obtain the *Total-step method* (*Jacobi method*):

$$x^{(k+1)} = D^{-1}(L + R)x^{(k)} + D^{-1}b \tag{4.1.6}$$

or in formula

$$x_j^{(k+1)} = \frac{1}{a_{jj}}(-\sum_{i \neq j} a_{ji}x_i^{(k)} + b_j),\ j = 1, \cdots, n,\ k = 0, 1, \cdots. \tag{4.1.7}$$

Example 4.1.2 If $D - L$ is nonsingular in (4.1.5), then the choices of $F = D - L$ and $G = R$ as in (4.1.2) are possible and yields the so-called *Single-step method* (*Gauss-Seidel method*):

$$x^{(k+1)} = (D - L)^{-1}Rx^{(k)} + (D - L)^{-1}b \tag{4.1.8}$$

or in formula

$$x_j^{(k+1)} = \frac{1}{a_{jj}}(-\sum_{i<j} a_{ji}x_i^{(k+1)} - \sum_{i>j} a_{ji}x_i^{(k)} + b_j),\ j = 1, \cdots, n,\ k = 1, 2, \cdots. \tag{4.1.9}$$

- Total-Step method=TSM=Jacobi method.
- Single-Step method=SSM=Gauss-Seidel method.

We now consider (4.1.1)-(4.1.4) once again:

Theorem 4.1.1 *Let $1 \notin \sigma(T)$ and x be the unique solution of (4.1.3). The sequence $x^{(k+1)} = Tx^{(k)} + f$ converges to x for arbitrary initial vector $x^{(0)}$ if and only if $\rho(T) < 1$*

Proof: We define the error

$$\varepsilon^{(k)} = x^{(k)} - x. \tag{4.1.10}$$

Then

$$\varepsilon^{(k)} = x^{(k)} - x = Tx^{(k-1)} - f - Tx + f = T\varepsilon^{(k-1)}$$

or

$$\varepsilon^{(k)} = T^k \varepsilon^{(0)}.$$

Theorem 1.2.4 shows that $\varepsilon^{(k)} \to 0$ if and only if $\rho(T) < 1$. ■

We now consider the following point of views on the Examples 4.1.1 and 4.1.2:

(i) flops counts per iteration step.

(ii) Convergence speed.

Let $\| \quad \|$ be a vector norm, and $\|T\|$ be the corresponding operator norm. Then

$$\frac{\|\varepsilon^{(m)}\|}{\|\varepsilon^{(0)}\|} = \frac{\|T^m \varepsilon^{(0)}\|}{\|\varepsilon^{(0)}\|} \leq \|T^m\|. \tag{4.1.11}$$

Here $\|T^m\|^{\frac{1}{m}}$ is a measure for the average diminution of error $\varepsilon^{(m)}$ per iteration step. We call

$$R_m(T) = -\ln(\|T^m\|^{\frac{1}{m}}) = -\frac{1}{m}\ln(\|T^m\|) \tag{4.1.12}$$

the average of convergence rate for m iterations.

The larger is $R_m(T)$, so the better is convergence rate. Let $\sigma = (\|\varepsilon^{(m)}\|/\|\varepsilon^{(0)}\|)^{\frac{1}{m}}$. From (4.1.11) and (4.1.12) we get

$$\sigma \leq \|T^m\|^{\frac{1}{m}} \leq e^{-R_m(T)},$$

or

$$\sigma^{1/R_m(T)} \leq \frac{1}{e}.$$

That is, after $1/R_m(T)$ steps in average the error is reduced by a factor of $1/e$. Since $R_m(T)$ is not easy to determine, we now consider $m \to \infty$. Since

$$\lim_{m\to\infty} \|T^m\|^{\frac{1}{m}} = \rho(T),$$

it follows

$$R_\infty(T) = \lim_{m\to\infty} R_m(T) = -\ln \rho(T). \tag{4.1.13}$$

R_∞ is called the asymptotic convergence rate. It holds always $R_m(T) \leq R_\infty(T)$.

Example 4.1.3 Consider the Dirichlet boundary-value problem (Model problem):

$$-\Delta u \equiv -u_{xx} - u_{yy} = f(x,y), \quad 0 < x, y < 1, \tag{4.1.14}$$

$$u(x,y) = 0 \ \ (x,y) \in \partial\Omega,$$

for the unit square $\Omega := \{x, y | 0 < x, y < 1\} \subseteq \mathbb{R}^2$ with boundary $\partial\Omega$.

To solve (4.1.14) by means of a difference methods, one replaces the differential operator by a difference operator. Let

$$\begin{aligned} \Omega_h &:= \{(x_i, y_i) | i, j = 1, \cdots, N+1\}, \\ \partial\Omega_h &:= \{(x_i, 0), (x_i, 1), (0, y_j), (1, y_j) | i, j = 0, 1, \cdots, N+1\}, \end{aligned}$$

where $x_i = ih,\ y_j = jh,\ i,j = 0,1,\cdots,N+1,\ h := \frac{1}{N+1},\ N \geq 1,$ an integer.

The differential operator $-u_{xx} - u_{yy}$ can be replaced for all $(x_i, y_i) \in \Omega_h$ by the difference operator:

$$\frac{4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} \tag{4.1.15}$$

up to an error $\tau_{i,j}$. Therefore for sufficiently small h one can thus expect that the solution $z_{i,j}$, for $i,j = 1,\cdots,N$ of the linear system

$$\begin{aligned} &4z_{i,j} - z_{i-1,j} - z_{i+1,j} - z_{i,j-1} - z_{i,j+1} = h^2 f_{i,j}, i,j = 1,\cdots,N, \\ &z_{0,j} = z_{N+1,j} = z_{i,0} = z_{i,N+1} = 0, i,j = 0,1,\cdots,N+1, \end{aligned} \tag{4.1.16}$$

obtained from (4.1.15) by omitting the error $\tau_{i,j}$, agrees approximately with the $u_{i,j}$. Let

$$z = [z_{1,1}, z_{2,1}, \cdots, z_{N,1}, z_{1,2}, \cdots, z_{N,2}, \cdots, z_{1,N}, \cdots, z_{N,N}]^T \tag{4.1.17a}$$

and

$$b = h^2[f_{1,1}, \cdots, f_{N,1}, f_{1,2}, \cdots, f_{N,2}, \cdots, f_{1,N}, \cdots, f_{N,N}]^T. \tag{4.1.17b}$$

Then (4.1.16) is equivalent to a linear system $Az = b$ with the $N^2 \times N^2$ matrix.

$$A = \left[\begin{array}{cccc|cccc|cccc|cccc} 4 & -1 & & & -1 & & & & & & & & & & & \\ -1 & \ddots & \ddots & & & \ddots & & & & & & & & & & \\ & & \ddots & -1 & & & \ddots & & & & & & & & & \\ & & -1 & 4 & & & & -1 & & & & & & & & \\ \hline -1 & & & & 4 & -1 & & & \ddots & & & & & & & \\ & \ddots & & & -1 & \ddots & \ddots & & & \ddots & & & & & & \\ & & \ddots & & & & \ddots & -1 & & & \ddots & & & & & \\ & & & -1 & & & -1 & 4 & & & & \ddots & & & & \\ \hline & & & & \ddots & & & & \ddots & & & & -1 & & & \\ & & & & & \ddots & & & & \ddots & & & & \ddots & & \\ & & & & & & \ddots & & & & \ddots & & & & \ddots & \\ & & & & & & & \ddots & & & & \ddots & & & & -1 \\ \hline & & & & & & & & -1 & & & & 4 & -1 & & \\ & & & & & & & & & \ddots & & & -1 & \ddots & \ddots & \\ & & & & & & & & & & \ddots & & & \ddots & -1 \\ & & & & & & & & & & & -1 & & & -1 & 4 \end{array}\right].$$

$$\equiv \begin{bmatrix} A_{1,1} & A_{1,2} & & \\ A_{2,1} & A_{2,2} & \ddots & \\ & \ddots & \ddots & A_{N-1,N} \\ & & A_{N,N-1} & A_{N,N} \end{bmatrix}. \tag{4.1.18}$$

Let $A = D - L - R$. The matrix $J = D^{-1}(L + R)$ belongs to the Jacobi method (TSM). The N^2 eigenvalues and eigenvectors of J can be determined explicitly. We can verify at

once, by substitution, that N^2 vectors $z^{(k,l)}$, $k,l = 1,\cdots,N$ with components

$$z_{i,j}^{(k,l)} := \sin\frac{k\pi i}{N+1}\sin\frac{l\pi j}{N+1},\ \ 1 \le i,j \le N,$$

satisfy

$$Jz^{(k,l)} = \lambda^{(k,l)} z^{(k,l)} \tag{4.1.19}$$

with

$$\lambda^{(k,l)} := \frac{1}{2}(\cos\frac{k\pi}{N+1} + \cos\frac{l\pi}{N+1}),\ \ 1 \le k,l \le N.$$

J thus has eigenvalues $\lambda^{(k,l)}$, $1 \le k,l \le N$. Then we have

$$\rho(J) = \lambda_{1,1} = \cos\frac{\pi}{N+1} = 1 - \frac{\pi^2 h^2}{2} + O(h^4) \tag{4.1.20}$$

and

$$R_\infty(J) = -\ln(1 - \frac{\pi^2 h^2}{2} + O(h^4)) = \frac{\pi^2 h^2}{2} + O(h^4). \tag{4.1.21}$$

These show that

(i) TSM converges; Nevertheless,

(ii) Diminution of h will not only enlarge the flop counts per step, but also the convergence speed will drastically make smaller.

sectionSome Remarks on nonnegative matrices

4.1.1 Some theorems and definitions

$\rho(T)$: A measure of quality for convergence.

Definition 4.1.1 *A real $m \times n$-matrix $A = (a_{ik})$ is called nonnegative (positive), denoted by $A \ge 0$ $(A > 0)$, if $a_{ik} \ge 0$ (> 0), $i = 1,\cdots,m$, $k = 1,\cdots,n$.*

Remark 4.1.1 *Let $K_n = \{x | x_i \ge 0, i = 1,\cdots,n\} \subseteq \mathbb{R}^n$. It holds*

$$A \in \mathbb{R}^{m\times n},\ A \ge 0 \Leftrightarrow AK_n \subset K_m.$$

Especially, for $m = n$, $A \ge 0 \Leftrightarrow AK \subset K, K = K$ is a cone.

Let $\tilde{N} = \{1, 2, \cdots, n\}$.

Definition 4.1.2 *An $m \times n$-matrix A is called reducible, if there is a subset $I \subset \tilde{N}$, $I \ne \phi$, $I \ne \tilde{N}$ such that $i \in I$, $j \notin I \Rightarrow a_{ij} = 0$. A is not reducible $\Leftrightarrow$ A is irreducible.*

Remark 4.1.2 *$G(A)$ is the directed graph associated with the matrix A. If A is an $n \times n$-matrix, then $G(A)$ consists of n vertices $P_1,\cdots,P_n$ and there is an (oriented) arc $P_i \to P_j$ in $G(A)$ precisely if $a_{ij} \ne 0$.*

It is easily shown that A is irreducible if and only if the graph $G(A)$ is connected in the sense that for each pair of vertices (P_i, P_j) in $G(A)$ there is an oriented path from P_i to P_j. i.e., if $i \neq j$, there is a sequence of indices $i = i_1, i_2, \cdots, i_s = j$ such that $(a_{i_1,i_2} \cdots a_{i_{s-1},i_s}) \neq 0$.

Lemma 4.1.1 *If $A \geq 0$ is an irreducible $n \times n$ matrix, then $(I + A)^{n-1} > 0$.*

Proof: It is sufficient to prove for any $x \geq 0$, $(I+A)^{n-1}x > 0$. Let $x_{k+1} = (I+A)x_k$ be a sequence of nonnegative vectors, for $0 \leq k \leq n-2$ with $x_0 = x$. We now verify that x_{k+1} has fewer zero components than does x_k for every $0 \leq k \leq n-2$. Since $x_{k+1} = x_k + Ax_k$, it is clear that x_{k+1} has no more zero components than x_k.

If x_{k+1} and x_k has exactly the same number of zero components, then for a suitable permutation P we have

$$Px_{k+1} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \quad Px_k = \begin{bmatrix} \beta \\ 0 \end{bmatrix}, \quad \alpha > 0, \ \beta > 0, \ \alpha, \beta \in \mathbb{R}^m, \ 1 \leq m \leq n.$$

Then

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

This implies $A_{21}\beta = 0$. But $A_{21} \geq 0$ and $\beta > 0$, it follows $A_{21} = 0$. It contradicts that A is irreducible. Thus x_{k+1} has fewer components and x_k has at most $(n-k-1)$ zero component. Hence

$$x_{n-1} = (I + A)^{n-1}x_0$$

is a positive vector. ■

(See also Miroslav Fiedler: "*Special Matrices and their applications in Numerical Mathematics* " for the following theorems.)

Lemma 4.1.2 *If A, B are squared matrices and $|A| \leq B$, then $\rho(A) \leq \rho(B)$. In particular, $\rho(A) \leq \rho(|A|)$.*

Proof: Suppose $|A| \leq B$, but $\rho(A) > \rho(B)$. Let s satisfy $\rho(A) > s > \rho(B)$, $P = (\frac{1}{s})A$ and $Q = (\frac{1}{s})B$. Then $\rho(P) = s^{-1}\rho(A) > 1$, $\rho(Q) = s^{-1}\rho(B) < 1$. This means that $\lim_{k\to\infty} Q^k = 0$. But $|P^k| \leq |P|^k \leq Q^k$ this implies $\lim_{k\to 0} P^k = 0$, i.e., $\rho(P) < 1$. Contradiction! ■

Lemma 4.1.3 *Let $A \geq 0$, $z \geq 0$. If ξ is a real number satisfies $Az > \xi z$, then $\rho(A) > \xi$.*

Proof: Assume $\xi \geq 0$. Clearly, $z \neq 0$. Since $Az > \xi z$, there is an $\varepsilon > 0$ such that $Az \geq (\xi + \varepsilon)z$. It means that $B = (\xi + \varepsilon)^{-1}A$ satisfies $Bz \geq z$. Thus,

$$B^k z \geq B^{k-1}z \geq \cdots \geq z, \text{ for } k > 0 \text{ (integer)}.$$

Hence B^k does not converge to the null matrix. This implies, $\rho(B) \geq 1$ and $\rho(A) \geq \xi + \varepsilon > \xi$. ■

Theorem 4.1.4 (Perron-Frobenius Theorem) *Let $A \geq 0$ be irreducible. Then $\rho(A)$ is a simple positive eigenvalue of A and there is a positive eigenvector belonging to $\rho(A)$. No nonnegative eigenvector belongs to any other eigenvalue of A.*

Remark 4.1.3 *$\rho(A)$ is called a Perron root of A. The eigenvector corresponding to $\rho(A)$ is called a Perron vector.*

Lemma 4.1.1 (Perron Lemma) *If $A > 0$, then $\rho(A)$ is a positive eigenvalue of A and there is only one linearly independent eigenvector corresponding to the eigenvalue $\rho(A)$. Moreover, this eigenvector may be chosen to be positive.*

Proof: The lemma holds for $n = 1$. Let $n > 1$ and $A > 0$. There exists an eigenvalue λ of A such that $\rho(A) = |\lambda|$. Let

$$Au = \lambda u, \quad u \neq 0. \tag{4.1.1}$$

Since $|\alpha v + \beta w| \leq \alpha|v| + \beta|w|$, if $v, w \in \mathbb{C}$, $\alpha,\ \beta \in \mathbb{R}_+$, then

$$\text{“=” holds} \Leftrightarrow \textit{exits} \text{ complex unit } \eta \text{ such that } \eta v \geq 0 \text{ and } \eta w \geq 0.$$

Generalization: Since $|\sum_{i=1}^{n} \alpha_i v_i| \leq \sum_{i=1}^{n} \alpha_i |v_i|$, for $v_1, \ldots, v_n \in \mathbb{C}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_+$. Then

$$\text{“=” holds} \Leftrightarrow \exists \text{ complex unit } \eta \text{ such that } \eta v_i \geq 0, \quad i = 1, \cdots, n.$$

Use this result to show u in (4.1.1) has the property that there is a complex unit η such that

$$\eta u_i \geq 0, \text{ for } i = 1, \cdots, n. \tag{4.1.2}$$

To prove this, assume (4.1.2) does not hold. Then we have

$$|\lambda||u_k| = |\sum_{j=1}^{n} a_{kj} u_j| < \sum_{j=1}^{n} a_{kj}|u_j|$$

in k-th equation of (4.1.1). By the above statement, this is true for $k = 1, \cdots, n$. Thus,

$$A|u| > |\lambda||u|.$$

From Lemma 4.1.3 follows that $|\lambda| < \rho(A)$, which contradicts that $|\lambda| = \rho(A)$.

Therefore, the inequality (4.1.2) implies $v = \eta u$, $v \neq 0$ nonnegative and from (4.1.1) follows

$$Av = \lambda v. \tag{4.1.3}$$

If $v_k \neq 0$ and thus $v_k > 0$, then the k-th equation in (4.1.3) gives $\lambda > 0$. Hence $\lambda = \rho(A)$ and using (4.1.3) again follows $v > 0$.

In particular, we have proved the implication: if λ is an eigenvalue such that $|\lambda| = \rho(A)$ and if u is an associated eigenvector then $|u| > 0$.

Suppose that there are two linearly independent eigenvectors $v = (v_i)$ and $w = (w_i)$, belonging to λ. As $v \neq 0$, there is an integer k such that $v_k \neq 0$. The vector $z = w - (w_k v_k^{-1})v$ is also an eigenvector of A belonging to λ. Since $z \neq 0$, but $z_k = 0$, this contradicts the proved results in above which states that $|z| > 0$. ■

Corollary 4.1.1 *Let $A > 0$.* Then $|\lambda| < \rho(A)$ for every eigenvalue $\lambda \neq \rho(A)$.

Proof: $|\lambda| \le \rho(A)$ for all eigenvalues λ of A. Suppose $|\lambda| = \rho(A)$ and $Ax = \lambda x,\ x \neq 0$. By Perron Lemma there is an $w = e^{-i\theta}x > 0$ for some $\theta \in \mathbb{R}$ such that $Aw = \lambda w$. But then $\lambda = \rho(A)$. Contradictions! ■

Proof of Theorem 4.1.4: Since $A \ge 0$ irreducible, $(I+A)^{n-1}$ is positive by Lemma 4.1.1. Also,

$$(I + A^T)^{n-1} = ((I + A)^{n-1})^T$$

is positive. By Perron Lemma there is an $y > 0$ such that

$$y^T(I + A)^{n-1} = \rho((I + A)^{n-1})y^T. \tag{4.1.4}$$

Let λ be the eigenvalue of A satisfying $|\lambda| = \rho(A)$ and $Ax = \lambda x$, $x \neq 0$. Further,

$$\rho^2(A)|x| \le \rho(A)A|x| = A\rho(A)|x| \le A^2|x|,$$

and in general

$$\rho^k(A)|x| \le A^k|x|, \text{ for } k = 1, 2, \cdots . \tag{4.1.5}$$

Hence

$$(1 + \rho(A))^{n-1}|x| \le (I + A)^{n-1}|x|. \tag{4.1.6}$$

Multiplying y^T from left it implies

$$(1 + \rho(A))^{n-1}(y^T|x|) \le y^T(I + A)^{n-1}|x|.$$

From (4.1.4) follows that

$$R.H.S = \rho((I + A)^{n-1})y^T|x|.$$

Since $y^T|x| > 0$, it implies

$$(1 + \rho(A))^{n-1} \le \rho((I + A)^{n-1}). \tag{4.1.7}$$

The eigenvalues of $(I + A)^{n-1}$ are of the form $(1 + \alpha)^{n-1}$, where α is an eigenvalue of A. Hence there is an eigenvalue μ of A such that

$$|(1 + \mu)^{n-1}| = \rho((I + A)^{n-1}). \tag{4.1.8}$$

On the other hand, we have $|\mu| \le \rho(A)$. Substituting into (4.1.7), we get

$$(1 + \rho(A))^{n-1} \le |(1 + \mu)^{n-1}|$$

and further

$$1 + \rho(A) \le |1 + \mu| \le 1 + |\mu| \le 1 + \rho(A).$$

Since the left-hand and right-hand sides coincide, we have equality everywhere. Thus $\mu \ge 0$ and hence $\mu = \rho(A)$.

Equality is valid in all the inequalities that we have added, i.e., in (4.1.5). For $k = 1$, it follows

$$A|x| = \rho(A)|x| \ \text{ or } \ A|x| = \mu|x|.$$

In view of (4.1.6) and (4.1.8) follows

$$(I + A)^{n-1}|x| = |1 + \mu|^{n-1}|x| = \rho((I + A)^{n-1})|x|.$$

Using Perron's Lemma, we get $|x| > 0$.

From this we know that there is only one linearly independent eigenvector belonging to eigenvalue μ by the same argument as that used in the last paragraph of the proof of Perron's Lemma. Moreover, $\rho(A) > 0$ as A is distinct from the null matrix $(n > 1)$!. Consequently, we want to claim: $\rho(A)$ is a simple eigenvalue of A if and only if

(i) there is a unique linearly independent eigenvector of A to λ, say u and also only one linearly independent eigenvector of A^T belonging to λ, say v.

(ii) $v^T u \neq 0$.

Indeed, only one linearly independent eigenvector of A, say u, belongs to $\rho(A)$. Moreover $u > 0$. Similarly, $A^T \geq 0$ irreducible. The respective eigenvector v of A^T (to $\rho(A)$) can be chosen positive as well $v > 0$. Therefore $v^T u > 0$ and by Schur Lemma follows that $\rho(A)$ is simple.

Finally, we show that no nonnegative eigenvector belongs to any other eigenvalue. Suppose $Az = \xi z,\ z \geq 0$ and $\xi \neq \rho(A)$. We have shown that A^T has a positive eigenvector, say $w > 0$. Then,

$$A^T w = \rho(A) w.$$

But,

$$w^T A z = w^T \xi z = \xi (w^T z),$$

i.e.,

$$w^T A z = \rho(A)(w^T z),$$

which is a contradiction in view of $\rho(A) - \xi \neq 0$ and $w^T z > 0$. ■

Theorem 4.1.5 *Let $A \geq 0, x > 0$. Define the quotients:*

$$q_i(x) \equiv \frac{(Ax)_i}{x_i} = \frac{1}{x_i} \sum_{k=1}^{n} a_{ik} x_k, \ \textit{for } i = 1, \cdots, n. \tag{4.1.9}$$

Then

$$\min_{1 \leq i \leq n} q_i(x) \leq \rho(A) \leq \max_{1 \leq i \leq n} q_i(x). \tag{4.1.10}$$

If A is irreducible, then it holds additionally, either

$$q_1 = q_2 = \cdots q_n \ (\textit{then } x = \mu z, \ \ q_i = \rho(A)) \tag{4.1.11}$$

or

$$\min_{1 \leq i \leq n} q_i(x) < \rho(A) < \max_{1 \leq i \leq n} q_i(x). \tag{4.1.12}$$

Proof: We first assume that A is irreducible. Then A^T is irreducible. From Theorem 4.1.4 there exists $y > 0$ such that $A^T y = \rho(A^T) y = \rho y$. Since $Ax = Qx$ with $Q = diag(q_1, \cdots, q_n)$, it follows

$$\sum_{i=1}^{n} q_i y_i x_i = y^T Q x = y^T A x = \rho y^T x = \rho \sum_{i=1}^{n} y_i x_i$$

or

$$\sum_{i=1}^{n}(q_i - \rho)y_i x_i = 0.$$

Now there is either $q_i - \rho = 0$, for all $i = 1, \cdots, n$, that is (4.1.11) holds or there is a $q_i \neq \rho$. Since $y_i x_i > 0$, so (4.1.12) holds. (4.1.10) follows from the consideration of the limiting case. ■

Theorem 4.1.6 *The statements in Theorem 4.1.5 can be formulated as: Let $A \geq 0, x > 0$. (4.1.10) corresponds:*

$$\begin{cases} Ax \leq \mu x \quad \Rightarrow \quad \rho \leq \mu, \\ \\ Ax \geq \nu x \quad \Rightarrow \quad \nu \leq \rho. \end{cases} \tag{4.1.13}$$

Let $A \geq 0$, irreducible, $x > 0$. (4.1.12) corresponds :

$$\begin{cases} Ax \leq \mu x, \ \ Ax \neq \mu x \quad \Rightarrow \quad \rho < \mu, \\ Ax \geq \nu x, \ \ Ax \neq \nu x \quad \Rightarrow \quad \nu < \rho. \end{cases} \tag{4.1.14}$$

Theorem 4.1.7 *(Perron and Frobenius 1907-1912, see Varga pp.30) Let $A \geq 0$ irreducible. Then*

(i) *$\rho = \rho(A)$ is a simple eigenvalue;*

(ii) *There is a positive eigenvector z associated to ρ, i.e., $Az = \rho z, z > 0$;*

(iii) *If $Ax = \lambda x$, $x \geq 0$, then $\lambda = \rho$, $x = \alpha z$, $\alpha > 0$. i.e, if x is any nonnegative eigenvector of A, then x is a multiplicity of z;*

(iv) *$A \leq B, A \neq B \Longrightarrow \rho(A) < \rho(B)$.*

Note that (i), (ii) and (iii) follows by Theorem 4.1.4 immediately. The proof of (iv) follows from Lemma 4.1.12 in Appendix. ■

Theorem 4.1.8 *(See Varga pp.46) If $A \geq 0$, then*

(i) *$\rho = \rho(A)$ is an eigenvalue.*

(ii) *There is a $z \geq 0$, $z \neq 0$ with $Az = \rho z$.*

(iii) *$A \leq B \Longrightarrow \rho(A) \leq \rho(B)$.*

Note that If $A \geq 0$ reducible, then A is a limit point of irreducible nonnegative matrices. Hence some parts of Theorem 4.1.7 are preserved.

Appendix

Let $A = [a_{ij}] \geq 0$ be irreducible and $x \geq 0$ be any vector. Let

$$r_x \equiv \min_{x_i > 0} \left\{ \frac{\sum_{j=1}^{n} a_{ij} x_j}{x_i} \right\} \geq 0.$$

Then,

$$r_x = \sup\{\rho \geq 0 \mid Ax \geq \rho x\}. \tag{4.1.15}$$

Consider

$$r = \sup_{x>0, x\neq 0} \{r_x\}. \tag{4.1.16}$$

Since r_x and $r_{\alpha x}$ have the same value for all $\alpha > 0$, we only consider $\|x\| = 1$ and $x \geq 0$. Let $P = \{x \mid x \geq 0,\ \|x\| = 1\}$ and $Q = \{y \mid (I + A)^{n-1} x,\ x \in P\}$. From Lemma 4.1.1 follows Q consists only of positive vector. Multiplying $Ax \geq r_x x$ by $(I + A)^{n-1}$, we get $Ay \geq r_x y$ (by (9.15)). Thus $r_y \geq r_x$.

The quantity r of (4.1.16) can be defined equivalently as

$$r = \sup_{y \in Q} \{r_y\}. \tag{4.1.17}$$

Note that r_y: $Q \to \mathbb{R}$ taking its maximum. As P is compact, so is Q, and as r_y is a continuous function on Q, there exists a positive z for which

$$Az \geq rz \tag{4.1.18}$$

and no vector $w \geq 0$ exists for which $Aw > rw$.

All non-negative nonzero z satifying (4.1.18) is called an extremal vector of the matrix A.

Lemma 4.1.9 *Let $A \geq 0$ be irreducible. The quantity r of (4.1.16) is positive. Moreover, each extremal vector z is a positive eigenvector of A with corresponding eigenvalue r. i.e., $Az = rz,\ z > 0$.*

Proof: If ξ is positive and $\xi_i = 1$, then since A is irreducible, no row of A can vanish. Thus no component of $A\xi$ can vanish. Thus $r_\xi > 0$. Proving that $r > 0$. Let z be an extremal vector which

$$Az - rz = \eta, \quad \eta \geq 0.$$

If $\eta \neq 0$, then some component of η is positive. Multiplying both sides by $(I + A)^{n-1}$ we get

$$Aw - rw > 0, \quad w = (I + A)^{n-1} z > 0.$$

Therefore, $r_w > r$ which contradicts (4.1.17). Thus $Az = rz$. Since $w > 0$ and $w = (1 + r)^{n-1} z$, it follows $z > 0$. ■

Lemma 4.1.10 *Let $A = [a_{ij}] \geq 0$ be irreducible and $B = [b_{ij}]$ be a complex matrix with $|B| \leq A$. If β is any eigenvalue of B, then*

$$|\beta| \leq r, \tag{4.1.19}$$

where r is the positive constant of (4.1.16). Moreover, equality in (4.1.19) holds, i.e., $\beta = re^{i\varphi}$, if and only if, $|B| = A$, and B has the form

$$B = e^{i\varphi} DAD^{-1}, \tag{4.1.20}$$

where D is diagonal whose diagonal entries have modulus unity.

Proof: If $By = \beta y, \;\; y \neq 0$, then $\beta y_i = \sum_{j=1}^{n} b_{ij} y_i, \;\; 1 \leq i \leq n$. Thus,

$$|\beta||y| \leq |B||y| \leq A|y|.$$

This implies, $|\beta| \leq r_{|y|} \leq r$. Hence, (4.1.19) is proved.

If $|\beta| = r$, then $|y|$ is an extremal vector of A. From Lemma 4.1.9 follows that $|y|$ is a positive eigenvector of A corresponding to the eigenvalue r. Thus,

$$r|y| = |B||y| = A|y|. \tag{4.1.21}$$

Since $|y| > 0$, from (4.1.21) and $|B| \leq A$ follows

$$|B| = A. \tag{4.1.22}$$

For vector y, ($|y| > 0$), we set

$$D = \text{diag}\left\{ \frac{y}{|y_1|}, \cdots, \frac{y_n}{|y_n|} \right\}.$$

Then

$$y = D|y| \tag{4.1.23}$$

Setting $\beta = re^{i\varphi}$, then $By = \beta y$ can be written as

$$C|y| = r|y| \tag{4.1.24}$$

where

$$C = e^{-i\varphi} D^{-1} BD. \tag{4.1.25}$$

From (4.1.21) and (4.1.24) follows that

$$C|y| = |B||y| = A|y|. \tag{4.1.26}$$

From the definition of C in (4.1.25) follows that $|C| = |B|$. Combining with (4.1.22) we have

$$|C| = |B| = A \tag{4.1.27}$$

Thus, from (4.1.26) we conclude that $C|y| = |C||y|$, and as $|y| > 0$, follows $C = |C|$, and thus $C = A$ from (4.1.27). Combing this result with (4.1.25) gives

$$B = e^{i\varphi} DAD^{-1}.$$

Conversely, it is obvious that B has the form in (4.1.20), then $|B| = A$. So, B has an eigenvalue β with $|\beta| = r$. ■

Setting $B = A$ in Lemma 4.1.10, we have

Corollary 4.1.11 *If $A \geq 0$ is irreducible, then the positive eigenvalue r of Lemma 4.1.9 equals the spectral radius $\rho(A)$ of A.*

Lemma 4.1.12 *If $A \geq 0$ is irreducible and B is any principal squared submatrix of A, then $\rho(B) < \rho(A)$.*

Proof: There is a permutation P such that

$$C = \begin{bmatrix} A_{11} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and } PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Clearly, $0 \leq C \leq PAP^T$ and $\rho(C) = \rho(B) = \rho(A_{11})$. But as $C = |C| \neq PAP^T$ follows that $\rho(B) < \rho(A)$. ■

4.1.2 The theorems of Stein-Rosenberg

Remark 4.1.4 Let D be nonsingular in the standard decomposition (4.1.5)

$$A = D - L - R.$$

Consider $\tilde{A} = D^{-1}A = \tilde{D} - \tilde{L} - \tilde{R}$, where $\tilde{D} = I$, $\tilde{L} = D^{-1}L$ and $\tilde{R} = D^{-1}R$. Then we have

$$D^{-1}(L + R) = D^{-1}L + D^{-1}R = \tilde{D}^{-1}(\tilde{L} + \tilde{R})$$

and

$$(D - L)^{-1}R = (I - D^{-1}L)^{-1}D^{-1}R = (\tilde{D} - \tilde{L})^{-1}\tilde{R}.$$

When we investigate TSM and SSM, we can without loss of generality suppose that $D = I$. Therefore in the following paragraph we assume that

$$A = I - L - R. \tag{4.1.28}$$

The iteration matrices of TSM and SSM become

$$J = L + R, \tag{4.1.29}$$

$$H = (I - L)^{-1}R, \tag{4.1.30}$$

respectively. If $L \geq 0$ and $R \geq 0$, then J and $H = (I - L)^{-1}R = (I + L + \cdots + L^{n-1})R$ are nonnegative. Here, we have $L^n = 0$.

Theorem 4.1.13 *Let $A = I - L - R$, $L \geq 0$, $R \geq 0$, $n \geq 2$. Then precisely one of the following relationships holds:*

(i) $0 = \rho(H) = \rho(J)$,

(ii) $0 < \rho(H) < \rho(J) < 1$,

(iii) $\rho(H) = \rho(J) = 1$,

(iv) $\rho(H) > \rho(J) > 1$.

Proof: We will only give the proof of the case when A is irreducible. Hence the case (i) does not occur. If A is reducible, then we can transform the reducible matrices into irreducible matrices by using the normalform method. The method is very skillful and behind our discussion, so we assume that A is irreducible.

(a) claim: $\rho(H) > 0$.

Let $z > 0$ be given. Then $b = (I - L)^{-1}Rz \geq 0$. Certainly $Rz \neq 0$, thus $b = Rz + LRz + \cdots + L^{n-1}Rz \neq 0$. Hence $I = \{i \mid b_i = 0\} \neq \{1, 2, \cdots, n\} = \tilde{N}$. Because $Rz = b - Lb$, for $i \in I$ we have

$$0 = b_i = \sum_{k>i} r_{ik}z_k + \sum_{k<i} l_{ik}b_k$$

and

$$\begin{cases} r_{ik} = 0,\ i \in I,\ k > i,\ k \notin I, \\ l_{ik} = 0,\ i \in I,\ k < i,\ k \notin I, \end{cases}$$

and $a_{ik} = 0$ for all $i \in I,\ k \notin I$. Since A is irreducible, it follows that $I = \emptyset$. For $b > 0$ and from Theorem 4.1.5 follows that $0 < \min\limits_{1 \leq i \leq n} \left\{ \frac{b_i}{z_i} \right\} \leq \rho(H)$.

(b) Let $x \geq 0$ be the eigenvector of H corresponding to $\rho_H = \rho(H)$ (by Theorem 4.1.8). Let $\rho_J = \rho(J)$. Since $(I - L)^{-1}Rx = \rho_H x$, thus

$$\frac{1}{\rho_H}Rx = x - Lx \ \text{ or } \ x = (L + \frac{1}{\rho_H}R)x.$$

Since A is irreducible, we can conclude that $L + \frac{1}{\rho_H}R$ is also irreducible. According to Theorem 4.1.7 (iii) we have

$$1 = \rho(L + \frac{1}{\rho_H}R) \tag{4.1.31}$$

and $x > 0$. Now we define the real value function

$$\phi(t) = \rho(L + \frac{1}{t}R), \quad t > 0. \tag{4.1.32}$$

From Theorem 4.1.7 (iv) we can conclude that $\phi(t)$ is strictly (monotonic) decreasing in t. On the other hand, $t\phi(t) = \rho(tL + R),\ \ t > 0$ is strictly (monotone) increasing in t.

(case 1) $\rho_H < 1$: Since $\rho_J = \phi(1)$, it implies that

$$\rho_J = \phi(1) = \rho(L + R) > \rho(\rho_H L + R) = \rho_H \rho(L + \frac{1}{\rho_H R}) = \rho_H. \quad \text{(by (4.1.31))}$$

(case2) $\rho_H = 1$: $\rho(L + R) = \rho_J = 1$.

(case 3) $\rho_H > 1$: $\rho_J = \phi(1) > \phi(\rho_H) = 1$ and

$$\rho_J = \phi(1) = \rho(L + R) < \rho(\rho_H L + R) = \rho_H \rho(L + \frac{1}{\rho_H}R) = \rho_H.$$

■

Theorem 4.1.14 *If the off-diagonal elements in A ($A = I - L - R$) are nonpositive, then SSM is convergent if and only if TSM is convergent. Furthermore, SSM is asymptotically faster.*

Proof: The result follows immediately from theorem 4.1.13 and (4.1.13). ■

4.1.3 Sufficient conditions for convergence of TSM and SSM

Definition 4.1.3 *A real matrix B is called an M-matrix if $b_{ij} \leq 0, i \neq j$ and B^{-1} exists with $B^{-1} \geq 0$.*

In the following theorems we give some important equivalent conditions of the $M-$ matrix.

Theorem 4.1.15 *Let B be a real matrix with $b_{ij} \leq 0$ for $i \neq j$. Then the following statements are equivalent.*

(i) *B is an $M-$matrix.*

(ii) *There exists a vector $v > 0$ so that $Bv > 0$.*

(iii) *B has a decomposition $B = sI - C$ with $C \geq 0$ and $\rho(C) < s$.*

(iv) *For each decomposition $B = D - C$ with $D = diag\,(d_i)$ and $C \geq 0$, it holds: $d_i > 0$, $i = 1, 2, \cdots, n$, and $\rho(D^{-1}C) < 1$.*

(v) *There is a decomposition $B = D - C$, with $D = diag(d_i)$ and $C \geq 0$ it holds: $d_i > 0, i = 1, 2, \cdots, n$ and $\rho(D^{-1}C) < 1$.*
Further, if B is irreducible, then (6) is equivalent to (1)-(5).

(vi) *There exists a vector $v > 0$ so that $Bv \geq 0, \neq 0$.*

Proof:

(i) ⇒ (ii) : Let $e = (1, \cdots, 1)^T$. Since $B^{-1} \geq 0$ is nonsingular it follows $v = B^{-1}e > 0$ and $Bv = B(B^{-1}e) = e > 0$.

(ii) ⇒ (iii) : Let $s > \max(b_{ii})$. It follows $B = sI - C$ with $C \geq 0$. There exists a $v > 0$ with $Bv = sv - Cv$ (via (ii)), also $sv > Cv$. From the statement (4.1.13) in Theorem 4.1.6 follows $\rho(C) < s$.

(iii) ⇒ (i) : $B = sI - C = s(I - \frac{1}{s}C)$. For $\rho(\frac{1}{s}C) < 1$ and from Theorem 1.2.6 follows that there exists a series expansion $(I - \frac{1}{s}C)^{-1} = \sum_{\nu=0}^{\infty} (\frac{1}{s}C)^k$. Since the terms in sum are nonnegative, we get $B^{-1} = \frac{1}{s}(I - \frac{1}{s}C)^{-1} \geq 0$.

(ii) ⇒ (iv) : From $Bv = Dv - Cv > 0$ follows $Dv > Cv \geq 0$ and $d_i > 0$, for $i = 1, 2, \cdots, n$. Hence $D^{-1} \geq 0$ and $v > D^{-1}Cv \geq 0$. From (4.1.13) follows that $\rho(D^{-1}C) < 1$.

(iv) ⇒ (v) : Trivial.

(v) ⇒ (i) : Since $\rho(D^{-1}C) < 1$, it follows from Theorem 1.2.6 that $(I - D^{-1}C)^{-1}$ exists and equals to $\sum_{k=0}^{\infty} (D^{-1}C)^k$. Since the terms in sum are nonnegative, we have $(I - D^{-1}C)^{-1}$ is nonnegative and $B^{-1} = (I - D^{-1}C)^{-1}D^{-1} \geq 0$.

(ii) ⇒ (vi) : Trivial.

(vi) ⇒ (v) : Consider the decomposition $B = D - C$, with $d_i = b_{ii}$. Let $\{I = i \mid d_i \leq 0\}$. From $d_i v_i - \sum_{k \neq i} c_{ik} v_k \geq 0$ follows $c_{ik} = 0$ for $i \in I$, and $k \neq i$. Since $Bv \geq 0, \neq 0 \Rightarrow I \neq \{1, \cdots, n\}$. But B is irreducible $\Rightarrow I = \emptyset$ and $d_i > 0$. Hence for $Dv >, \neq Cv$ also $v >, \neq D^{-1}Cv$ and (4.1.14) show that $\rho(D^{-1}C) < 1$. ■

Remark 4.1.5 *Theorem 4.1.15 can also be described as follows: If $a_{ij} \leq 0$, $i \neq j$, then TSM and SSM converge if and only if A is an M-matrix.*

Proof: By (i) ⇔ (iv) and (i) ⇔ (v) of the previous theorem and Theorem 4.1.14. ■

Lemma 4.1.16 *Let A be an arbitrary complex matrix and define $|A| = [|a_{ij}|]$. If $|A| \leq C$, then $\rho(A) \leq \rho(C)$. Especially $\rho(A) \leq \rho(|A|)$.*

Proof: There is a $x \neq 0$ with $Ax = \lambda x$ and $|\lambda| = \rho(A)$. Hence

$$\rho(A)|x_i| = |\sum_{k=1}^{n} a_{ik} x_k| \leq \sum_{k=1}^{n} |a_{ik}||x_k| \leq \sum_{k=1}^{n} c_{ik}|x_k|.$$

Thus,

$$\rho(A)|x| \leq C|x|.$$

If $|x| > 0$, then from (4.1.13) we have $\rho(A) \leq \rho(C)$. Otherwise, let $I = \{i \mid x_i \neq 0\}$ and C_I be the matrix, which consists of the ith row and ith column of C with $i \in I$. Then we have $\rho(A)|x_I| \leq C_I|x_I|$. Here $|x_I|$ consists of ith component of $|x|$ with $i \in I$. Then from $|x_I| > 0$ and (4.1.13) follows $\rho(A) \leq \rho(C_I)$. We now fill C_I with zero up to an $n \times n$ matrix $\tilde{C}_I$. Then $\tilde{C}_I \leq C$. Thus, $\rho(C_I) = \rho(\tilde{C}_I) \leq \rho(C)$ (by Theorem 4.1.8(iii)). ■

Theorem 4.1.17 *Let A be an arbitrary complex matrix. It satisfies*
either (Strong Row Sum Criterion)*:*

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|, \quad i = 1, \cdots, n. \tag{4.1.33}$$

or (Weak Row Sum Criterion)*:*

$$\begin{aligned} \sum_{j \neq i} |a_{ij}| &\leq |a_{ii}|, \ i = 1, \cdots, n, \\ &< |a_{i_0 i_0}|, \ \textit{at least one } i_0, \end{aligned} \tag{4.1.34}$$

for A irreducible. Then TSM and SSM both are convergent.

Proof: Let $A = D - L - R$. From (4.1.33) and (4.1.34) D must be nonsingular and then as in Remark 4.1.4 we can w.l.o.g. assume that $D = I$. Now let $B = I - |L| - |R|$. Then (4.1.33) can be written as $Be > 0$. From Theorem 4.1.15(ii) and (i) follows that B is an M-matrix.

(4.1.34) can be written as $Be \geq 0$, $Be \neq 0$. Since A is irreducible, also B, from Theorem 4.1.15 (vi) and (i) follows that B is an M-matrix.

Especially, from theorem 4.1.15(i), (iv) and Theorem 4.1.13 follows that

$$\rho(|L| + |R|) < 1 \text{ and } \rho((I - |L|)^{-1}|R|) < 1.$$

Now Lemma 4.1.16 shows that

$$\rho(L + R) \leq \rho(|L| + |R|) < 1.$$

So TSM is convergent. Similarly,

$$\begin{aligned} \rho((I - L)^{-1}R) &= \rho(R + LR + \cdots + L^{n-1}R) \\ &\leq \rho(|R| + |L||R| + \cdots + |L|^{n-1}|R|) \\ &= \rho((I - |L|)^{-1}|R|) < 1. \end{aligned}$$

So SSM is convergent. ■

4.2 Relaxation Methods (Successive Over-Relaxation (SOR) Method)

Consider the standard decomposition (4.1.5)

$$A = D - L - R$$

for solving the linear system (4.1.1) $Ax = b$. The single-step method (SSM)

$$(D - L)x^{i+1} = Rx^{(i)} + b$$

can be written in the form

$$x^{(i+1)} = x^{(i)} + \{D^{-1}Lx^{(i+1)} + D^{-1}Rx^{(i)} + D^{-1}b - x^{(i)}\} := x^{(i)} + v^{(i)}. \tag{4.2.1}$$

Consider a general form of (4.2.1)

$$x^{(i+1)} = x^{(i)} + \omega v^{(i)} \tag{4.2.2}$$

with constant ω. Also (4.2.2) can be written as

$$Dx^{(i+1)} = Dx^{(i)} + \omega Lx^{(i+1)} + \omega Rx^{(i)} + \omega b - \omega Dx^{(i)}.$$

Then

$$x^{(i+1)} = (D - \omega L)^{-1}((1 - \omega)D + \omega R)x^{(i)} + \omega(D - \omega L)^{-1}b. \tag{4.2.3}$$

We now assume that $D = I$ as above. Then (4.2.3) becomes

$$x^{(i+1)} = (I - \omega L)^{-1}((1 - \omega)I + \omega R)x^{(i)} + \omega(I - \omega L)^{-1}b \tag{4.2.4}$$

with the iteration matrix

$$L_\omega := (I - \omega L)^{-1}((1 - \omega)I + \omega R). \tag{4.2.5}$$

These methods is called for

$\omega < 1$: under relaxation,
$\omega = 1$: single-step method,
$\omega > 1$: over relaxation. (In general: relaxation methods.)

We now try to choose an ω such that $\rho(L_\omega)$ is possibly small. But this is only under some special assumptions possible. we first list a few qualitative results about $\rho(L_\omega)$.

Theorem 4.2.1 *Let $A = D - L - L^*$ be hermitian and positive definite. Then the relaxation method is convergent for $0 < \omega < 2$.*

Proof: We claim that each eigenvalue of L_ω has absolute value smaller than 1 (i.e., $\rho(L_\omega) < 1$). Let $\lambda \in \sigma(L_\omega)$. Then there is an $x \neq 0$ with

$$\lambda(D - \omega L)x = ((1-\omega)D + \omega L^*)x. \tag{4.2.6}$$

It holds obviously

$$\begin{aligned} 2\lambda(D - \omega L) &= \lambda((2-\omega)D + \omega(D - 2L)) \\ &= \lambda((2-\omega)D + \omega A + \omega(L^* - L)) \end{aligned}$$

and

$$\begin{aligned} 2[(1-\omega)D + \omega L^*] &= (2-\omega)D + \omega(-D + 2L^*) \\ &= (2-\omega)D - \omega A + \omega(L^* - L). \end{aligned}$$

Hence multiplying (4.2.6) by x^* we get

$$\begin{aligned} &\lambda((2-\omega)x^*Dx + \omega x^*Ax + \omega x^*(L^* - L)x) \\ &= (2-\omega)x^*Dx - \omega x^*Ax + \omega x^*(L^* - L)x \end{aligned}$$

or by $d = x^*Dx > 0$, $a := x^*Ax > 0$ and $x^*(L^* - L)x := is$, $s \in R$ we get

$$\lambda((2-\omega)d + \omega a + i\omega s) = (2-\omega)d - \omega a + i\omega s.$$

Dividing above equation by ω and setting $\mu = (2-\omega)/\omega$, we get

$$\lambda\{\mu d + a + is\} = \mu d - a + is.$$

For $0 < \omega < 2$ we have $\mu > 0$ and $\mu d + is$ is in the right half plane. Therefore the distance from a to $\mu d + is$ is smaller than that from $-a$. So we have $|\lambda| = |\frac{\mu d + is - a}{\mu d + is + a}| < 1$. ■

Theorem 4.2.2 *Let A be Hermitian and nonsingular with positive diagonal. If SSM converges, then A is positive definite.*

Proof: Let $A = D - L - L^*$. For any matrix C it holds:

$$\begin{aligned} A - (I - AC^*)A(I - CA) &= A - A + ACA + AC^*A - AC^*ACA \\ &= AC^*(C^{*-1} + C^{-1} - A)CA. \end{aligned}$$

For special case that $C = (D - L)^{-1}$ we have

$$C^{*-1} + C^{-1} - A = D - L^* + D - L - D + L + L^* = D$$

and

$$I - CA = (D - L)^{-1}(D - L - A) = (D - L)^{-1}L^* = H.$$

Hence we obtain

$$A - H^*AH = A(D - L)^{-*}D(D - L)^{-1}A =: B$$

Thus $H^*AH = A - B$. D is positive definite, obviously so is B (since $(D-L)^{-1}A$ is nonsingular). Because $\rho(H) < 1$, for any $\varepsilon_0 \in C^n$ the sequence $\{\varepsilon_m\}_{m=1}^{\infty}$ defined by $\varepsilon_m := H^m\varepsilon_0$ converges to zero. Therefore the sequence $\{\varepsilon_m^* A\varepsilon_m\}_{m=1}^{\infty}$ also converges to zero. Furthermore, we have

$$\varepsilon_{m+1}^* A\varepsilon_{m+1} = \varepsilon_m^* H^* AH\varepsilon_m = \varepsilon_m^* A\varepsilon_m - \varepsilon_m^* B\varepsilon_m < \varepsilon_m^* A\varepsilon_m, \tag{4.2.7}$$

because $B > 0$ is positive definite. If A is not positive definite, then there is a $\varepsilon_0 \in \mathbb{C}^n \backslash \{0\}$ with $\varepsilon_0^* A\varepsilon_0 \leq 0$. This is a contradiction that $\{\varepsilon_m^* A\varepsilon_m\} \to 0$ and (4.2.7). ■

4.2.1 Determination of the Optimal Parameter ω for 2-consistly Ordered Matrices

For an important class of matrices the more qualitative assertions of Theorems 4.2.1 and 4.2.2 can be considerably sharpened. This is the class of consistly ordered matrices. The optimal parameter ω_b with

$$\rho(L_{\omega_b}) = \min_{\omega} \rho(L_\omega)$$

can be determined. We consider $A = I - L - R$.

Definition 4.2.1 *A is called 2-consistly ordered, if the eigenvalues of $\alpha L + \alpha^{-1}R$ are independent of α.*

Example 4.2.1 $A = -\begin{bmatrix} 0 & R \\ L & 0 \end{bmatrix} + I,$

$$\alpha L + \alpha^{-1}R = \begin{bmatrix} 0 & \alpha^{-1}R \\ \alpha L & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \alpha I \end{bmatrix} \begin{bmatrix} 0 & R \\ L & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \alpha^{-1}I \end{bmatrix}.$$

This shows that $\alpha L + \alpha^{-1}R$ is similar to $L + R$, so the eigenvalues are independent to α. A is 2-consistently ordered. ■

Let $A = I - L - R$, $J = L + R$. Let $s_1^{(i)}$, $s_2^{(i)}$, ... denote the lengths of all closed oriented path (oriented cycles)

$$P_i \to P_{k_1} \to P_{k_2} \to \cdots \to P_{k_s^{(i)}} = P_i$$

in $G(J)$ which leads from P_i to P_i. Denoting by l_i the greatest common divisor: $l_i = g.c.d.(s_1^{(i)}, s_2^{(i)}, \cdots)$.

Definition 4.2.2 *The Graph $G(J)$ is called 2-cyclic if $l_1 = l_2 = \cdots = l_n = 2$ and weakly 2-cyclic if all l_i are even.*

Definition 4.2.3 *The matrix A has property A if there exists a permutation P such that $PAP^T = \begin{bmatrix} D_1 & M_1 \\ M_2 & D_2 \end{bmatrix}$ with D_1 and D_2 diagonal.*

Theorem 4.2.3 *For every $n\times n$ matrix A with property A and $a_{ii}\neq 0$, $i=1,\cdots,n$, there exists a permutation P such that $\bar{A}=D(I-L-R)$ of the permuted matrix $\bar{A}:=PAP^T$ is 2-consistly ordered.*

Proof: There is a permutation P such that

$$PAP^T=\begin{bmatrix} D_1 & M_1 \\ M_2 & D_2 \end{bmatrix}=D(I-L-R)$$

with

$$D=\begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix},\quad L=-\begin{bmatrix} 0 & 0 \\ D_2^{-1}M_2 & 0 \end{bmatrix} \text{ and } R=-\begin{bmatrix} 0 & D_1^{-1}M_1 \\ 0 & 0 \end{bmatrix}.$$

For $\alpha\neq 0$, we have

$$\begin{aligned} J(\alpha) &= -\begin{bmatrix} 0 & \alpha^{-1}D_1^{-1}M_1 \\ \alpha D_2^{-1}M_2 & 0 \end{bmatrix}=-S_\alpha\begin{bmatrix} 0 & D_1^{-1}M_1 \\ D_2^{-1}M_2 & 0 \end{bmatrix}S_\alpha^{-1} \\ &= S_\alpha J(1)S_\alpha^{-1}, \end{aligned}$$

where $S_\alpha:=\begin{bmatrix} I_1 & 0 \\ 0 & \alpha I_2 \end{bmatrix}$. ■

Theorem 4.2.4 *An irreducible matrix A has property A if and only if $G(J)$ is weakly 2-cyclic. (Without proof!)*

Example 4.2.2 Block tridiagonal matrices

$$A=\begin{bmatrix} D_1 & A_{12} & & \\ A_{21} & D_2 & \ddots & \\ & \ddots & \ddots & A_{N-1,N} \\ & & A_{N,N-1} & D_N \end{bmatrix}.$$

If all D_i are nonsingular, then

$$J(\alpha)=\begin{bmatrix} 0 & \alpha^{-1}D_1^{-1}A_{12} & \cdots & 0 \\ \alpha D_2^{-1}A_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha^{-1}D_{N-1}^{-1}A_{N-1,N} \\ 0 & \cdots & \alpha D_N^{-1}A_{N,N-1} & 0 \end{bmatrix},$$

which obeys the relation $J(\alpha)=S_\alpha J(1)S_\alpha^{-1}$, with $S_\alpha=\mathrm{diag}\{I_1,\alpha I_2,\cdots,\alpha^{N-1}I_N\}$. Thus A is 2-consistly ordered. ■

The other description: $G(L+R)$ is bipartite.

Example 4.2.3 $A=\begin{bmatrix} 1 & b_1 & & 0 \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & c_{n-1} & 1 \end{bmatrix}$ is 2-consistly ordered. The eigenvalues are the roots of $\det(A-\lambda I)=0$. The coefficients of above equation appear only those products b_ic_i. For $\alpha L+\alpha^{-1}R$, we substitute b_i and c_i by $\frac{1}{\alpha}b_i$ and αc_i, respectively, then the products are still b_ic_i. Therefore eigenvalues are independent of α. ■

Examples 4.2.1 and 4.2.2 are 2-cyclic.

Example 4.2.4

$$A = \begin{bmatrix} 1 & a & b \\ 0 & 1 & 0 \\ c & d & 1 \end{bmatrix}, \quad -\alpha L - \alpha^{-1} R = \begin{bmatrix} 0 & \alpha^{-1}a & \alpha^{-1}b \\ 0 & 0 & 0 \\ \alpha c & \alpha d & 0 \end{bmatrix}.$$

The coefficients of characteristic polynomial are independent to α, so A is 2-consistly ordered. But $G(L+R)$ is not bipartite, so not 2-cyclic. ■

If A is 2-consistly ordered, then $L+R$ and $-(L+R)$ ($\alpha = -1$) has the same eigenvalues. The nonzero eigenvalues of $L+R$ appear in pairs. Hence

$$\det(\lambda I - L - R) = \lambda^m \prod_{i=1}^{r} (\lambda^2 - \mu_i^2), \; n = 2r + m \; (m = 0, \text{ possible}). \tag{4.2.8}$$

Theorem 4.2.5 *Let A be 2-consistly ordered, $a_{ii} = 1$, $\omega \neq 0$. Then hold:*

(i) *If $\lambda \neq 0$ is an eigenvalue of L_ω and μ satisfies the equation*

$$(\lambda + \omega - 1)^2 = \lambda \mu^2 \omega^2, \tag{4.2.9}$$

then μ is an eigenvalue of $L+R$ (so is $-\mu$).

(ii) *If μ is an eigenvalue of $L+R$ and λ satisfies the equation (4.2.9), then λ is an eigenvalue of L_ω.*

Remark 4.2.1 *If $\omega = 1$, then $\lambda = \mu^2$, and $\rho((I-L)^{-1}R) = (\rho(L+R))^2$.*

Proof: We first prove the identity

$$\det(\lambda I - sL - rR) = \det(\lambda I - \sqrt{sr}(L+R)). \tag{4.2.10}$$

Since both sides are polynomials of the form $\lambda^n + \cdots$ and

$$sL + rR = \sqrt{sr}(\sqrt{\frac{s}{r}}L + \sqrt{\frac{r}{s}}R) = \sqrt{sr}(\alpha L + \alpha^{-1}R),$$

if $sr \neq 0$, then $sL + rR$ and $\sqrt{sr}(L+R)$ have the same eigenvalues. It is obviously also for the case $sr = 0$. The both polynomials in (4.2.10) have the same roots, so they are identical. For

$$\begin{aligned} \det(I - \omega L)\det(\lambda I - L_\omega) &= \det(\lambda(I - \omega L) - (1-\omega)I - \omega R) \\ &= \det((\lambda + \omega - 1)I - \omega\lambda L - wR) = \Phi(\lambda) \end{aligned}$$

and $\det(I - \omega L) \neq 0$, λ is an eigenvalue of L_ω if and only if $\Phi(\lambda) = 0$. From (4.2.10) follows

$$\Phi(\lambda) = \det((\lambda + \omega - 1)I - \omega\sqrt{\lambda}(L+R))$$

and that is (from (4.2.8))

$$\Phi(\lambda) = (\lambda + \omega - 1)^m \prod_{i=1}^{r} ((\lambda + \omega - 1)^2 - \omega^2 \lambda \mu_i^2), \tag{4.2.11}$$

where μ_i is an eigenvalue of $L + R$. Therefore, if μ is an eigenvalue of $(L + R)$ and λ satisfies (4.2.9), so is $\Phi(\lambda) = 0$, then λ is eigenvalue of L_ω. This shows (b).

Now if $\lambda \neq 0$ an eigenvalue of L_ω, then one factor in (4.2.11) must be zero. Let μ satisfy (4.2.9). Then

(i) $\mu \neq 0$: From (4.2.9) follows $\lambda + \omega - 1 \neq 0$, so

$$\begin{aligned} (\lambda + \omega - 1)^2 &= \lambda \omega^2 \mu_i^2, \quad \text{for one } i \text{ (from (4.2.11))}, \\ &= \lambda \omega^2 \mu^2, \quad \text{(from (4.2.9))}. \end{aligned}$$

This shows that $\mu = \pm \mu_i$, so μ is an eigenvalue of $L + R$.

(ii) $\mu = 0$: We have $\lambda + \omega - 1 = 0$ and

$$0 = \Phi(\lambda) = \det((\lambda + \omega - 1)I - \omega\sqrt{\lambda}(L + R)) = \det(-\omega\sqrt{\lambda}(L + R)),$$

i.e., $L + R$ is singular, so $\mu = 0$ is eigenvalue of $L + R$. ■

Theorem 4.2.6 *Let $A = I - L - R$ be 2-consistly ordered. If $L + R$ has only real eigenvalues and satisfies $\rho(L + R) < 1$, then it holds*

$$\rho(L_{\omega_b}) = \omega_b - 1 < \rho(L_\omega), \ \textit{for } \omega \neq \omega_b, \tag{4.2.12}$$

where

$$\omega_b = \frac{2}{1 + \sqrt{1 - \rho^2(L + R)}} \ \textit{(solve } \omega_b \textit{ in (4.2.9))}.$$

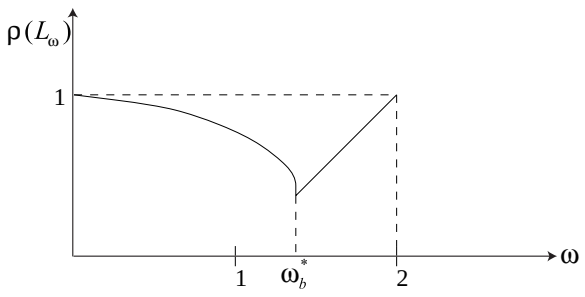


Figure 4.1: figure of $\rho(L_{\omega_b})$

One has in general,

$$\rho(L_\omega) = \begin{cases} \omega - 1, \ \textit{for } \omega_b \leq \omega \leq 2 \\ 1 - \omega + \frac{1}{2}\omega^2\mu^2 + \omega\mu\sqrt{1 - \omega + \frac{1}{4}\omega^2\mu^2}, \ \textit{for } 0 < \omega \leq \omega_b \end{cases} \tag{4.2.13}$$

Remark: We first prove the following Theorem proposed by Kahan: For arbitrary matrices A it holds

$$\rho(L_\omega) \geq |\omega - 1|, \text{ for all } \omega. \tag{4.2.14}$$

Proof: Since $\det(I - \omega L) = 1$ for all ω, the characteristic polynomial $\Phi(\lambda)$ of L_ω is

$$\begin{aligned} \Phi(\lambda) &= \det(\lambda I - L_\omega) = \det((I - \omega L)(\lambda I - L_\omega)) \\ &= \det((\lambda + \omega - 1)I - \omega\lambda L - \omega R). \end{aligned}$$

For $\prod_{i=1}^{n} \lambda_i(L_\omega) = \Phi(0) = \det((\omega - 1)I - \omega R) = (\omega - 1)^n$, it follows immediately that

$$\rho(L_\omega) = \max_i |\lambda_i(L_\omega)| \geq |\omega - 1|.$$

Proof of Theorem 4.2.6: By assumption the eigenvalues μ_i of $L + R$ are real and $-\rho(L + R) \leq \mu_i \leq \rho(L + R) < 1$. For a fixed $\omega \in (0, 2)$ (by (4.2.14) in the Remark it suffices to consider the interval (0,2)) and for each μ_i there are two eigenvalues $\lambda_i^{(1)}(\omega, \mu_i)$ and $\lambda_i^{(2)}(\omega, \mu_i)$ of L_ω, which are obtained by solving the quadratic equation (4.2.9) in λ. Geometrically, $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$ are obtained as abscissae of the points of intersection of the straight line $g_\omega(\lambda) = \frac{\lambda+\omega-1}{\omega}$ and the parabola $m_i(\lambda) := \pm\sqrt{\lambda}\mu_i$ (see Figure 4.2). The line $g_\omega(\lambda)$ has the slope $1/\omega$ and passes through the point (1,1). If $g_\omega(\lambda) \cap m_i(\lambda) = \phi$, then $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$ are conjugate complex with modulus $|\omega - 1|$ (from (4.2.9)). Evidently

$$\rho(L_\omega) = \max_i(|\lambda_i^{(1)}(\omega)|, |\lambda_i^{(2)}(\omega)|) = \max(|\lambda^{(1)}(\omega)|, |\lambda^{(2)}(\omega)|),$$

where $\lambda^{(1)}(\omega)$, $\lambda^{(2)}(\omega)$ being obtained by intersecting $g_\omega(\lambda)$ with $m(\lambda) := \pm\sqrt{\lambda}\mu$, with $\mu = \rho(L+R) = \max_i |\mu_i|$. By solving (4.2.9) with $\mu = \rho(L+R)$ for λ, one verifies (4.2.13) immediately, and thus also the remaining assertions of the theorem. ■

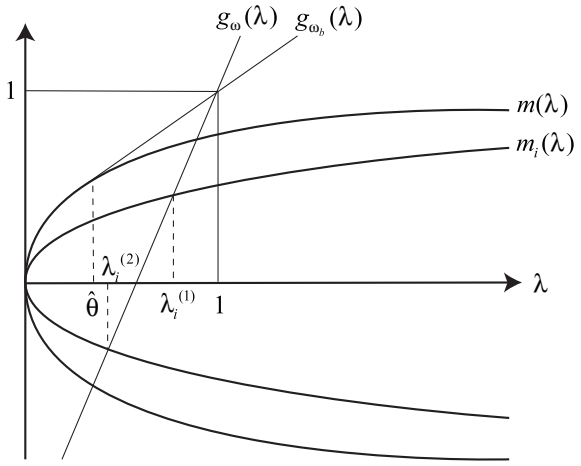


Figure 4.2: Geometrical view of $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$.

4.2.2 Practical Determination of Relaxation Parameter ω_b

For $\omega \in [1, \omega_b]$, from (4.2.13) in Theorem 4.2.6 we have

$$\rho(L_\omega) = [\frac{\omega\mu}{2} + \frac{1}{2}\sqrt{\mu^2\omega^2 - 4(\omega-1)}]^2 \tag{4.2.15}$$

or

$$\mu = \frac{\rho(L_\omega) + \omega - 1}{\omega\sqrt{\rho(L_\omega)}}. \tag{4.2.16}$$

Here $\mu := \rho(L+R)$. If μ is simple, then $\rho(L_\omega)$ is also a simple eigenvalue (See the proof of Theorem 4.2.6). So one can determine an approximation for $\rho(L_\omega)$ using power method (see later for details!): Let $\{x^{(k)}\}_{k=1}^{\infty}$ be the sequence of iterates, which generated by (4.2.6) with parameter ω. Let $e^{(k)} = x^{(k)} - x$ be the error vector which satisfies the relation $e^{(k)} = L_\omega^k e^{(0)}$ (Here $Ax = b$). We define $d^{(k)} := x^{(k+1)} - x^{(k)}$, for $k \in \mathbb{N}$. Then we have

$$\begin{aligned} x^{(k+1)} - x^{(k)} &= e^{(k+1)} - e^{(k)} = (L_\omega - I)e^{(k)} = (L_\omega - I)L_\omega^k e^{(0)} \\ &= L_\omega^k(L_\omega - I)e^{(0)} = L_\omega^k d^{(0)}. \end{aligned}$$

Hence $d^{(k)} = L_\omega^k d^{(0)}$. For sufficiently large $k \in \mathbb{N}$ we compute

$$q_k := \max_{1\le i\le n} \frac{|x_i^{(k+1)} - x_i^{(k)}|}{|x_i^{(k)} - x_i^{(k-1)}|}, \tag{4.2.17}$$

which is a good approximation for $\rho(L_\omega)$. We also determine the corresponding approximation for μ by (4.2.16) and the corresponding optimal parameter $\tilde{\omega}$ as (by Theorem 4.2.4):

$$\tilde{\omega} = 2/(1 + [1 - (q_k + \omega - 1)^2/(\omega^2 q_k)]^{1/2}). \tag{4.2.18}$$

4.2.3 Break-off Criterion for SOR Method

From $d^{(k)} = (L_\omega - I)e^{(k)}$ follows (for $\rho(L_\omega) < 1$) that $e^{(k)} = (L_\omega - I)^{-1}d^{(k)}$ and then

$$\|e^{(k)}\|_\infty \le \frac{1}{1 - \rho(L_\omega)}\|d^{(k)}\|_\infty.$$

With an estimate $q < 1$ for $\rho(L_\omega)$ one obtains the break-off criterion for a given $\varepsilon \in \mathbb{R}_+$

$$\begin{aligned} \|d^{(k)}\|_\infty &\le (1-q)\varepsilon, \quad \text{(for absolute error)}, \\ \|d(k)\|_\infty/\|x^{(k+1)}\|_\infty &\le (1-q)\varepsilon, \quad \text{(for relative error)}. \end{aligned}$$

The estimate q_k in (4.2.17) for the spectral radius $\rho(L_\omega)$ of SOR method is theoretically justified, if $\omega \le \omega_b$. But during the computation we cannot guarantee that the new $\tilde{\omega}$ also satisfies $\tilde{\omega} \le \omega_b$. Then an oscillation of q_k at $\tilde{\omega}$ may occur, and $1 - q_k$ can be considerably larger than $1 - \rho(L_{\tilde{\omega}})$; the break-off criterion may be satisfied too early. It is better to take $q := \max(q_k, \tilde{\omega} - 1)$ instead of q_k.

~~**Algorithm 4.2.1 (Successive Over-Relaxation Method)** *Let $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Let $A = D - L - R$ with D nonsingular. Suppose that A is 2-consistly ordered, all eigenvalues of $J := D^{-1}(L + R)$ are real and $\rho(J)$ is a simple eigenvalue of J satisfying $\rho(J) < 1$. [We apply a simple strategy to the following Algorithm, to perform a new updating after p iterative steps ($p \approx 5$).]*~~

Step 1: Choose a bound for machine precision $\varepsilon \in \mathbb{R}_+$ and a positive integer p, and a initial vector $x^{(0)} \in \mathbb{R}^n$. Let $\omega := 1$, $q := 1$ and $k := 0$.

Step 2: (Iterative step):
Compute for $i = 1, \ldots, n$,

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

If k is not positive integral multiplicity of p, then go to **Step 4**.

Step 3: (Adaptation of the Estimate of the Optimal Parameter):
Compute

$$q := \max_{1\le i\le n} \frac{|x_i^{(k+1)} - x_i^{(k)}|}{|x_i^{(k)} - x_i^{(k-1)}|}.$$

If $q > 1$, then go to **Step 5**.
Let $q := \max(q, \omega - 1)$ and $\omega := \frac{2}{1+\sqrt{1-\frac{(q+\omega-1)^2}{\omega^2 q}}}$.

Step 4: (Break-off criterion): If

$$\frac{\max\limits_{1\le i\le n}\left|x_i^{(k+1)} - x_i^{(k)}\right|}{\max\limits_{1\le i\le n}\left|x_i^{(k+1)}\right|} \le \varepsilon(1-q),$$

then stop.

Step 5: Let $k := k + 1$ and go to step 2.

4.3 Application to Finite Difference Methods: Model Problem (Example 4.1.3)

We consider the Dirichlet boundary-value problem (Model problem) as in Example 8.3. We shall solve a linear system $Az = b$ of the $N^2 \times N^2$ matrix A as in (4.1.18).
To Jacobi method: The iterative matrix is

$$J = L + R = \frac{1}{4}(4I - A).$$

Graph $G(J)$ ($N = 3$) is connected and weakly 2-cyclic. Thus, A is irreducible and has property A. It is easily seen that A is 2-consistly ordered (Exercise!).
To Gauss-Seidel method: The iterative matrix is

$$H = (I - L)^{-1}R.$$

From the Remark of Theorem 4.2.5 and (4.1.20) follows that

$$\rho(H) = \rho(J)^2 = \cos^2 \frac{\pi}{N+1}.$$

According to Theorem 4.2.6 the optimal relaxation parameter ω_b and $\rho(L_{\omega_b})$ are given by

$$\omega_b = \frac{2}{1 + \sqrt{1 - \cos^2 \frac{\pi}{N+1}}} = \frac{2}{1 + \sin \frac{\pi}{N+1}} \tag{3.1}$$

and

$$\rho(L_{\omega_b}) = \frac{\cos^2 \frac{\pi}{N+1}}{(1 + \sin \frac{\pi}{N+1})^2}. \tag{3.2}$$

The number $k = k(N)$ with $\rho(J)^k = \rho(L_{\omega_b})$ indicates that the k steps of Jacobi method produce the same reduction as one step of the optimal relaxation method. Clearly,

$$k = \ln \rho(L_{\omega_b}) / \ln \rho(J). \tag{3.3}$$

Now for small z one has $\ln(1+z) = z - z^2/2 + O(z^3)$ and for large N we have

$$\cos\left(\frac{\pi}{N+1}\right) = 1 - \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^4}).$$

Thus that

$$\ln \rho(J) = \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^4}).$$

Similarly,

$$\begin{aligned} \ln \rho(L_{\omega_b}) &= 2[\ln \rho(J) - \ln(1 + \sin \frac{\pi}{N+1})] \\ &= 2[-\frac{\pi^2}{2(N+1)^2} - \frac{\pi}{N+1} + \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^3})] \\ &= -\frac{2\pi}{N+1} + O(\frac{1}{N^3}) \text{ (for large } N). \end{aligned}$$

and

$$k = k(N) \approx \frac{4(N+1)}{\pi}. \tag{3.4}$$

The optimal relaxation method is more than N times as fast as the Jacobi method. The quantities

$$R_J := \frac{-\ln 10}{\ln \rho(J)} \approx 0.467(N+1)^2. \tag{3.5}$$

$$R_H := \frac{1}{2} R_J \approx 0.234(N+1)^2 \tag{3.6}$$

$$R_{L_{\omega_b}} := -\frac{\ln 10}{\ln \rho(L_{\omega_b})} \approx 0.367(N+1) \tag{3.7}$$

indicate the number of iterations required in the Jacobi, the Gauss-Seidel method, and the optimal relaxation method, respectively, in order to reduce the error by a factor of 1/10.

4.4 Block Iterative Methods

A natural block structure

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{N1} & \cdots & A_{NN} \end{bmatrix},$$

where A_{ii} are square matrices. In addition, if all A_{ii} are nonsingular, we introduce block iterative methods relative to the given partition π of A, which is analogous to (4.1.5):

$$A = D_\pi - L_\pi - R_\pi$$

with

$$D_\pi := \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & A_{NN} \end{bmatrix}, \tag{4.1a}$$

$$L_\pi := -\begin{bmatrix} 0 & 0 & \cdots & 0 \\ A_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ A_{N1} & \cdots & A_{N,N-1} & 0 \end{bmatrix},$$

$$R_\pi := -\begin{bmatrix} 0 & A_{12} & \cdots & A_{1N} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & A_{N-1,N} \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{4.1b}$$

One obtains the block Jacobi method (block total-step method) for the solution of $Ax = b$ by choosing in (4.1.4) analogously to (4.1.6) or (4.1.7), $F := D_\pi$. One thus obtains

$$D_\pi x^{(i+1)} = b + (L_\pi + R_\pi)x^{(i)} \tag{4.2}$$

or

$$A_{jj}x_j^{(i+1)} = b_j - \sum_{k \neq j} A_{jk}x_k^{(i)}, \text{ for } j = 1, \ldots, N, \ i = 0, 1, 2, \cdots. \tag{4.3}$$

We must solve system of linear equations of the form $A_{jj}z = y$, $j = 1, \cdots, N$. By the methods of Chapter 2, a triangular factorization (or a Cholesky factorization, etc.) $A_{jj} = L_j R_j$ we can reduce $A_{jj}z = y$ to the two triangular systems

$$L_j u = y \text{ and } R_j z = u.$$

For the matrix A in Example 8.3 (model problem): Here A_{jj} are positive definite tridiagonal $N \times N$ matrices.

$$A_{jj} = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}, \quad L_j = \begin{bmatrix} \times & 0 & \cdots & 0 \\ \times & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & \times & \times \end{bmatrix}.$$

The rate of convergence of (4.3) is determined by $\rho(J_\pi)$ of the matrix

$$J_\pi := L_\pi + R_\pi$$

with $L_\pi := D_\pi^{-1} L_\pi$ and $R_\pi := D_\pi^{-1} R_\pi$.

One can analogously to (4.1.8) define a block Gauss-Seidel method (block single-step method):

$$H_\pi := (I - L_\pi)^{-1} R_\pi$$

or

$$A_{jj} x_j^{(i+1)} = b_j - \sum_{k<j} A_{jk} x_k^{(i+1)} - \sum_{k>i} A_{jk} x_k^{(i)}, \text{ for } j = 1, \cdots, N, \ i = 0, 1, 2, \cdots . \tag{4.4}$$

As in Section 10, one can also introduce block relaxation methods through the choice

$$L_\omega^\pi = (I - \omega L_\pi)^{-1} [(1 - \omega) I + \omega R_\pi] \tag{4.5}$$

and

$$x^{(i+1)} = (I - \omega L_\pi)^{-1} ((1 - \omega) I + \omega R_\pi) x^{(i)} + \omega (I - \omega L_\pi)^{-1} b. \tag{4.6}$$

If one defines A as 2-consistly ordered whenever the eigenvalues of the matrices $J_\pi(\alpha) = \alpha L_\pi + \alpha^{-1} R_\pi$ are independent of α. Optimal relaxation factors are determined as in Theorem 4.2.6 with the help of $\rho(J_\pi)$. For the model problem (Example 8.3), relative to the partition given in (8.18), $\rho(J_\pi)$ can again be determined explicitly. One finds

$$\rho(J_\pi) = \frac{\cos \frac{\pi}{N+1}}{2 - \cos \frac{\pi}{N+1}} < \rho(J). \tag{4.7}$$

For the corresponding optimal block relaxation method one has asymptotically for $N \to \infty$,

$$\rho(L_{\omega_b}^\pi) \approx \rho(L_{w_b})^k$$

with $k = \sqrt{2}$ (Exercise!). The number of iterations is reduced by a factor $\sqrt{2}$ compared to the ordinary optimal relaxation method.

4.5 The ADI method of Peaceman and Rachford

4.5.1 ADI method (alternating-direction implicit iterative method)

Slightly generalizing the model problem (4.1.14), we consider the Poisson problem

$$\begin{cases} -u_{xx} - u_{yy} + \sigma u(x, y) = f(x, y), \text{ for } (x, y) \in \Omega, \\ \\ u(x, y) = 0, \text{ for } (x, y) \in \partial\Omega, \end{cases} \tag{4.5.1}$$

where $\Omega = \{(x, y) \mid 0 < x < 1, \ 0 < y < 1\} \subseteq \mathbb{R}^2$ with boundary $\partial\Omega$. Here $\sigma > 0$ is a constant and $f : \Omega \cup \partial\Omega \to \mathbb{R}$ continous function. Using the same discretization and the same notation as in Example 8.3, one obtains

$$4 z_{ij} - z_{i-1,j} - z_{i+1,j} - z_{i,j-1} - z_{i,j+1} + \sigma h^2 z_{ij} = h^2 f_{ij}, \ 1 \le i, \ j \le N \tag{4.5.2}$$

with $z_{0j} = z_{N+1,j} = z_{i,0} = z_{i,N+1} = 0$, $0 \le i, j \le N+1$ for the approximate values z_{ij} of $u_{ij} = u(x_i, y_j)$. To the decomposition

$$\begin{aligned} 4z_{ij} &- z_{i-1,j} - z_{i+1,j} - z_{i,j-1} - z_{i,j+1} + \sigma h^2 z_{ij} \\ &\equiv (2z_{ij} - z_{i-1,j} - z_{i+1,j}) + (2z_{ij} - z_{i,j-1} - z_{i,j+1}) + \left(\sigma h^2 z_{ij}\right), \end{aligned} \tag{4.5.3}$$

there corresponds a decomposition of the matrix A if the system $Az = b$, of the form $A = H + V + \Sigma$ (H: Horizontal, V: Vertical). Here H, V, Σ are defined by

$$w_{ij} = 2z_{ij} - z_{i-1,j} - z_{i+1,j}, \text{ if } w = Hz, \tag{4.5.4a}$$

$$w_{ij} = 2z_{ij} - z_{i,j-1} - z_{i,j+1}, \text{ if } w = Vz, \tag{4.5.4b}$$

$$w_{ij} = \sigma h^2 z_{ij}, \text{ if } w = \Sigma z. \tag{4.5.4c}$$

Σ is a diagonal matrix with nonnegative elements, H and V are both symmetric and positive definite, where $H = [\cdot]$ and $V = [\cdot]$. $A = H + V + \Sigma$ is now transformed equivalently into

$$(H + \frac{1}{2}\Sigma + rI)z = (rI - V - \frac{1}{2}\Sigma)z + b$$

and also

$$(V + \frac{1}{2}\Sigma + rI)z = (rI - H - \frac{1}{2}\Sigma)z + b.$$

Here r is an arbitrary real number. Let $H_1 := H + \frac{1}{2}\Sigma$, $V_1 := V + \frac{1}{2}\Sigma$, one obtains ADI method:

$$(H_1 + r_{i+1}I)z^{(i+1/2)} = (r_{i+1}I - V_1)z^{(i)} + b, \tag{4.5.5}$$

$$(V_1 + r_{i+1}I)z^{(i+1)} = (r_{i+1}I - H_1)z^{(i+1/2)} + b. \tag{4.5.6}$$

With suitable ordering of the variables z_{ij}, the matrices $H_1 + r_{i+1}I$ and $V_1 + r_{i+1}I$ are positive definite tridiagonal matrices (assuming $r_{i+1} \ge 0$), so that the systems (4.5.5) and (4.5.6) can easily be solved for $z^{(i+1/2)}$ and $z^{(i+1)}$ via a Cholesky factorization. Eliminating $z^{(i+1/2)}$ in (4.5.5) and (4.5.6) we get

$$z^{(i+1)} = T_{r_{i+1}} z^{(i)} + g_{r_{i+1}}(b) \tag{4.5.7}$$

with

$$T_r := (V_1 + rI)^{-1}(rI - H_1)(H_1 + rI)^{-1}(rI - V_1), \tag{4.5.8}$$

$$g_r(b) := (V_1 + rI)^{-1}[I + (rI - H_1)(H_1 + rI)^{-1}]b. \tag{4.5.9}$$

For the error $f_i := z^{(i)} - z$ it follows from (4.5.7) and the relation $z = T_{r_{i+1}} z + g_{r_{i+1}}(b)$ by subtraction, that

$$f_{i+1} = T_{r_{i+1}} f_i, \tag{4.5.10}$$

and therefore

$$f_m = T_{r_m} T_{r_{m-1}} \cdots T_{r_1} f_0. \tag{4.5.11}$$

In view of (4.5.10) and (4.5.11), r_i are to be determined so that the spectral radius $\rho(T_{r_m}, \cdots, T_{r_1})$ becomes as small as possible.

For the case $r_i = r$:

Theorem 4.5.1 *Under the assumption that H_1 and V_1 are positive definite, one has $\rho(T_r) < 1$, for all $r > 0$.*

Proof: V_1 and H_1 are positive definite. Therefore $(V_1 + rI)^{-1}$, and $(H_1 + rI)^{-1}$ exist, for $r > 0$, and hence also T_r of (4.5.8). The matrix

$$\begin{aligned} \tilde{T}_r &:= (V_1 + rI)T_r(V_1 + rI)^{-1} \\ &= [(rI - H_1)(H_1 + rI)^{-1}][rI - V_1)(V_1 + rI)^{-1}] \end{aligned}$$

is similar to T_r. Hence $\rho(T_r) = \rho(\tilde{T}_r)$. The matrix $\tilde{H} := (rI - H_1)(H_1 + rI)^{-1}$ has the eigenvalues $(r - \lambda_j)/(r + \lambda_j)$, where $\lambda_j = \lambda_j(H_1)$ are the eigenvalues of H_1. Since $r > 0,\ \lambda_j > 0$ it follows that $|(r - \lambda_j)/(r + \lambda_j)| < 1$ and thus $\rho(\tilde{H}) < 1$. Since H_1 also $\tilde{H}$ are symmetric, we have

$$\|\tilde{H}\|_2 = \rho(\tilde{H}) < 1.$$

In the same way one has

$$\|\tilde{V}\|_2 < 1.$$

Let $\tilde{V} := (rI - V_1)(V_1 + rI)^{-1}$. Thus

$$\rho(\tilde{T}_r) \le \|\tilde{T}_r\|_2 \le \|\tilde{H}\|_2 \|\tilde{V}\|_2 < 1.$$

The eigenvalues of T_r can be exhibited by

$$H_1 z^{(k,l)} = \mu_k z^{(k,l)}, \tag{4.5.12a}$$
$$V_1 z^{(k,l)} = \mu_l z^{(k,l)}, \tag{4.5.12b}$$
$$T_r z^{(k,l)} = \mu^{(k,l)} z^{(k,l)}, \tag{4.5.12c}$$

where $z_{ij}^{(k,l)} := \sin\frac{k\pi i}{N+1}\sin\frac{l\pi j}{N+1},\ 1 \le i, j \le N$, with

$$\mu^{(k,l)} = \frac{(r - \mu_l)(r - \mu_k)}{(r + \mu_l)(r + \mu_k)},\ \mu_j := 4\sin^2\frac{j\pi}{2(N+1)}, \tag{4.5.13}$$

so that

$$\rho(T_r) = \max_{1 \le j \le N} \left| \frac{r - \mu_j}{r + \mu_j} \right|^2.$$

One finally finds a result (Exercise!):

$$\min_{r>0} \rho(T_r) = \rho(L_{\omega_b}) = \frac{\cos^2\left(\frac{\pi}{N+1}\right)}{(1 + \sin\left(\frac{\pi}{N+1}\right))^2},$$

where ω_b characterizes the best (ordinary) relaxation method. The best ADI method assuming constant choice of parameters, has the same rate of convergence for the model problem as the optimal ordinary relaxation method. Since the individual iteration step in ADI method is a great deal more expensive than in the relaxation method, the ADI method would appear to be inferior. This is certainly true if $r = r_1 = r_2 = \cdots$ is chosen.
For the case $r_i \neq r$:

However, if one makes use of the option to choose a separate parameter r_i in each step, the picture changes in favor of the ADI method. Indeed

$$T_{r_i} \cdots T_{r_1} z^{(k,l)} = \mu^{(k,l)}_{r_i \cdots r_1} z^{(k,l)},$$

where

$$\mu^{(k,l)}_{r_i \cdots r_1} = \prod_{j=1}^{i} \frac{(r_j - \mu_l)(r_j - \mu_k)}{(r_j + \mu_l)(r_j + \mu_k)}.$$

Choosing $r_j := \mu_j$, for $j = 1, \cdots, N$, we have $\mu^{(k,l)}_{r_N, \cdots, r_1} = 0$, for $1 \leq k, l \leq N$, so that by the linear independence of the $z^{(k,l)}$, $T_{r_N} \cdots T_{r_1} = 0$. With this special choice of the r_j, the ADI method for the model problem terminates after N steps with the exact solution. This is a happy coincidence, which is due to the following essential assumptions:

(1) H_1 and V_1 have in common a set of eigenvectors which span the whole space.

(2) The eigenvalues of H_1 and V_1 are known.

Theorem 4.5.2 *For Two Hermitian matrices H_1 and $V_1 \in \mathbb{C}^{n \times n}$, there exist n linearly independent (orthogonal) vectors $z_1, \cdots, z_n$, which are common eigenvectors of H_1 and V_1,*

$$H_1 z_i = \sigma_i z_i,\ V_1 z_i = \tau_i z_i,\ \textit{for } i = 1, \cdots, n, \tag{4.5.14}$$

if and only if H_1 commutes with V_1, i.e., $H_1 V_1 = V_1 H_1$.

Proof: "⇒": From (4.5.14) it follows that

$$H_1 V_1 z_i = \sigma_i \tau_i z_i = V_1 H_1 z_i,\ \text{for } i = 1, 2, \cdots, n.$$

Since the z_i form a basis in $\mathbb{C}^n$, it follows at once that $H_1 V_1 = V_1 H_1$.
"⇐": Let $H_1 V_1 = V_1 H_1$. Let $\lambda_1 < \cdots < \lambda_r$ be the eigenvalues of V_1 with the multiplicities $\sigma(\lambda_i),\ i = 1, \cdots, r$. According to Theorem 1.1.1 there exists a unitary matrix U with

$$\Lambda_V := U^* V_1 U = \begin{bmatrix} \lambda_1 I_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r I_r \end{bmatrix}.$$

From $H_1 V_1 = V_1 H_1$ it follows immediately that $\tilde{H}_1 = \Lambda_V \tilde{H}_1$, with $\tilde{H}_1 := U^* H_1 U$. We partition $\tilde{H}_1$ analogously to Λ_V:

$$\tilde{H}_1 = \begin{bmatrix} H_{11} & \cdots & H_{1r} \\ \vdots & & \vdots \\ H_{r1} & \cdots & H_{rr} \end{bmatrix}.$$

By multiplying out

$$\tilde{H}_1 \Lambda_V = \Lambda_V \tilde{H}_1,$$

one obtains $H_{ij} = 0$, for $i \neq j$, since $\lambda_i \neq \lambda_j$. The H_{ii} are Hermitian of order $\sigma(\lambda_i)$. There exist unitary matrices $\bar{U}_i$ such that $\bar{U}_i^* H_{ii} \bar{U}_i = \Lambda_i$ (diagonal). For $\bar{U} = \begin{bmatrix} \bar{U}_1 & & \\ & \ddots & \\ & & \bar{U}_r \end{bmatrix} \in$

$\mathbb{C}^{n\times n}$, since $H_{ij}=0$, for $i\neq j$, it follows the relations

$$(U\bar{U})^*H_1(U\bar{U})=\bar{U}^*\tilde{H}_1\bar{U}=\Lambda_H=\begin{bmatrix}\Lambda_1 & & \\ & \ddots & \\ & & \Lambda_r\end{bmatrix},$$

$$\text{i.e.,}\quad H_1(U\bar{U})=(U\bar{U})\Lambda_H,$$

and

$$(U\bar{U})^*V_1(U\bar{U})=\bar{U}^*\Lambda_V\bar{U}=\Lambda_V$$

$$\text{i.e.,}\quad V_1(U\bar{U})=(U\bar{U})\Lambda_V,$$

so that $z_i:=(U\bar{U})e_i$ can be taken as n common orthogonal eigenvectors of H_1 and V_1. ■

We now assume in the following discussion that H_1 and V_1 are two positive definite commuting $n\times n$ matrices with (4.5.14) and that two numbers α, β are given such that $0<\alpha\leq\sigma_i,\ \tau_i\leq\beta$, for $i=1,\cdots,n$. Then

$$T_rz_i=\frac{(r-\sigma_i)(r-\tau_i)}{(r+\sigma_i)(r+\tau_i)}z_i,\ \text{for } r>0,\ i=1,2,\cdots,n.$$

gives the problem:

$$\begin{aligned}\rho(T_{r_m},\cdots,T_{r_1}) &= \max_{1\leq i\leq n}\prod_{j=1}^{m}\left|\frac{(r_j-\sigma_i)(r_j-\tau_i)}{(r_j+\sigma_i)(r_j+\tau_i)}\right| \\ &\leq \max_{\alpha\leq x\leq\beta}\prod_{j=1}^{m}\left|\frac{r_j-x}{r_j+x}\right|^2. \end{aligned}\tag{4.5.15}$$

For a given m, it is natural to choose $r_i>0$, $i=1,\cdots,m$, so that the function

$$\varphi(r_1,\cdots,r_m):=\max_{\alpha\leq x\leq\beta}\prod_{j=1}^{m}\left|\frac{r_j-x}{r_j+x}\right|,\tag{4.5.16}$$

becomes as small as possible. For each m it can be shown that there are uniquely determined number $\bar{r}_i$ with $\alpha<\bar{r}_i<\beta$, $i=1,\cdots,m$, such that

$$d_m(\alpha,\beta):=\varphi(\bar{r}_1,\ldots,\bar{r}_m)=\min_{r_i>0,\,1\leq i\leq m}\varphi(r_1,\ldots,r_m).\tag{4.5.17}$$

The optimal parameter $\bar{r}_1,\cdots,\bar{r}_m$ can even be given explicitly, for each m, in term of elliptic functions [see Young (1971) pp.518-525]. In the special case $m=2^k$, the relevant results will now be presented without proof [see Young (1971), Varga (1962)]. Let $r_i^{(m)},i=1,2\cdots,m$, denote the optimal ADI parameters for $m=2^k$. The $r_i^{(m)}$ and $d_m(\alpha,\beta)$ can be computed recursively by means of Gauss's arithmetic-geometric mean algorithm. It can be shown that

$$d_{2n}(\alpha,\beta)=d_n(\sqrt{\alpha\beta},\frac{\alpha+\beta}{2}).\tag{4.5.18}$$

The optimal parameter of the minimax problem (4.5.17), $r_i^{(2n)}$ and $r_i^{(n)}$, being related by

$$r_i^{(n)} = \frac{r_i^{(2n)} + \alpha\beta/r_i^{(2n)}}{2},\ i = 1, 2, \cdots, n. \tag{4.5.19}$$

Define $\alpha_0 := \alpha$, $\beta_0 := \beta$. Then

$$\alpha_{j+1} := \sqrt{\alpha_j \beta_j},\ \beta_{j+1} := \frac{\alpha_j + \beta_j}{2},\ j = 0, 1, \cdots, k-1. \tag{4.5.20}$$

Thus

$$\begin{aligned} d_{2^k}(\alpha_0, \beta_0) &= d_{2^{k-1}}(\alpha_1, \beta_1) = \cdots \\ &= d_1(\alpha_k, \beta_k) = \frac{\sqrt{\beta_k} - \sqrt{\alpha_k}}{\sqrt{\beta_k} + \sqrt{\alpha_k}}. \text{ (Exercise!)} \end{aligned} \tag{4.5.21}$$

The solution of $d_1(\alpha_k, \beta_k)$ can be found with $r_1^{(1)} = \sqrt{\alpha_k \beta_k}$. The optimal ADI parameter $r_i^{(m)}$, $i = 1, \cdots, m = 2^k$ can be computed as follows:

(i) $s_1^{(0)} := \sqrt{\alpha_k \beta_k}$.

(ii) For $j = 0, 1, \cdots, k-1$, determine $s_i^{(j+1)}, i = 1, 2, \cdots, 2^{j+1}$ as the 2^{j+1} solutions of the 2^j quadratic equations in x,

$$s_i^{(j)} = \frac{1}{2}(x + \frac{\alpha_{k-1-j}\beta_{k-1-j}}{x}),\ i = 1, 2, \cdots, 2^j. \tag{4.5.22}$$

(iii) Put $r_i^{(m)} := s_i^{(k)},\ i = 1, 2, \ldots, m = 2^k$.

The $s_i^{(j)}, i = 1, 2, \cdots, 2^j$ are just the optimal ADI parameters for the interval $[\alpha_{k-j}, \beta_{k-j}]$. Let us use these formulas to study the model problem (8.14)(8.16), with $m = 2^k$ fixed, and the asymptotic behavior of $d_{2^k}(\alpha, \beta)$ as $N \to \infty$. For α and β we take the best possible bounds

$$\alpha = 4\sin^2\frac{\pi}{2(N+1)},\ \beta = 4\sin^2\frac{N\pi}{2(N+1)} = 4\cos^2\frac{\pi}{2(N+1)}. \tag{4.5.23}$$

We then have

$$d_m(\alpha, \beta) \sim 1 - 4\sqrt[m]{\frac{\pi}{4(N+1)}} \tag{4.5.24}$$

as $N \to \infty$, $m := 2^k$.

Proof of (4.5.24): By induction on k. Let $c_k := \sqrt{\alpha_k/\beta_k}$. One obtains from ((4.5.20) and (4.5.21) that

$$d_{2^k}(\alpha, \beta) = \frac{1 - c_k}{1 + c_k} \tag{4.5.25}$$

and

$$c_{k+1}^2 = \frac{2c_k}{1 + c_k^2}. \tag{4.5.26}$$

In order to prove (4.5.24), it suffices to show that

$$c_k \sim 2\sqrt[2^k]{\frac{\pi}{4(N+1)}},\ N \to \infty. \tag{4.5.27}$$

It follows then from (4.5.25) that for $N \to \infty$, $d_{2^k}(\alpha,\beta) \sim 1 - 2c_k$. But (4.5.27) is true for $k = 0$, by using

$$c_0 = \tan\frac{\pi}{2(N+1)} \sim \frac{\pi}{2(N+1)}.$$

Thus, if (4.5.27) is valid for some $k \geq 0$, then it is also valid for $k+1$, because from (4.5.26) we have at once $c_{k+1} \sim \sqrt{2c_k}$, as $N \to \infty$. ■

In practice, the parameter r_i are often repeated cyclically, i.e., one chooses a fixed m ($m = 2^k$), then determines approximately the optimal ADI parameter $r_i^{(m)}$ belonging to this m, and finally takes for the ADI method the parameters

$$r_{jm+i} := r_i^{(m)} \text{ for } i = 1, 2, \cdots, m,\ j = 0, 1, \cdots.$$

If m individual steps of the ADI method are considered a "big iteration step", then the quantity

$$\frac{-\ln 10}{\ln \rho(Tr_m, \ldots, T_{r_1})}$$

indicates how many big iteration steps are required to reduce the error by a factor of $1/10$, i.e.,

$$R_{ADI}^{(m)} = -m\frac{\ln 10}{\ln \rho(T_{r_m}, \ldots, T_{r_1})}$$

indicates how many ordinary ADI steps, on the average, are required for the same purpose. In case of the model problem one obtains for the optimal choice of parameter and $m = 2^k$, by virtue of (4.5.15) and (4.5.24),

$$\rho(T_{r_m}, \ldots, T_{r_1}) \leq d_m(\alpha,\beta)^2 \sim 1 - 8\sqrt[m]{\frac{\pi}{4(N+1)}},\ N \to \infty,$$

$$\ln \rho(T_{r_m}, \ldots, T_{r_1}) \leq -8\sqrt[m]{\frac{\pi}{4(N+1)}},\ N \to \infty,$$

so that

$$R_{ADI}^{(m)} \leq \frac{m}{8}\ln(10)\sqrt[m]{\frac{4(N+1)}{\pi}},\ N \to \infty. \tag{4.5.28}$$

Comparing to (3.5)-(3.7) shows that for $m > 1$ the ADI method converges considerably faster than the optimal ordinary relaxation method. This convergence behavior establishes the practical significance of the ADI method.

4.5.2 The algorithm of Buneman for the solution of the discretized Poisson Equation

Consider the possion problem

$$\begin{cases} -u_{xx} - u_{yy} + \sigma u = f(x,y), & \text{for } (x,y) \in \Omega, \\ u(x,y) = 0, & \text{for } (x,y) \in \partial\Omega, \end{cases} \tag{4.5.29}$$

where $\Omega \equiv \{(x,y) \mid 0 < x < a,\ 0 < y < b\} \subseteq \mathbb{R}^2$, $\sigma > 0$ is a constant and $f : \Omega \cup \partial\Omega \to \mathbb{R}$ is a continuous function.

Discretizing (4.5.29): for the approximate z_{ij} of $u(x_i, y_j)$, $x_i = i\delta x$, $y_j = j\delta y$, $\delta x \equiv a/(p+1)$, $\delta y \equiv b/(q+1)$. We obtain the equation:

$$\frac{-z_{i-1,j} + 2z_{i,j} - z_{i+1,j}}{\delta x^2} + \frac{-z_{i,j-1} + 2z_{i,j} - z_{i,j+1}}{\delta y^2} + \sigma z_{i,j} = f_{ij} = f(x_i, y_j), \quad (4.5.30)$$

for $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, q$. Together with the boundary values

$$z_{0,j} \equiv z_{p+1,j} \equiv 0, \text{ for } j = 0, 1, \ldots, q+1,$$
$$z_{i,0} \equiv z_{i,q+1} \equiv 0, \text{ for } i = 0, 1, \ldots, p+1.$$

Let $z = [z_1^T, z_2^T, \cdots, z_q^T]^T$, $z_j = [z_{1j}, z_{2j}, \cdots, z_{pj}]^T$. Then (4.5.30) can be written in the forms

$$Mz = b \quad (4.5.31)$$

with

$$M = \begin{bmatrix} A & I & & \\ I & A & \ddots & \\ & \ddots & \ddots & I \\ & & I & A \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{bmatrix}, \quad (4.5.32)$$

where $I = I_p$, A is a $p \times p$ Hermitian tridiagonal matrix, and M consists of q block rows and columns.

We describe here only Buneman algorithm (1969). For related method see also Hockney (1969) and Swarztranber (1977). Now, (4.5.32) can be written as:

$$\begin{cases} Az_1 + z_2 = b_1, \\ z_{j-1} + Az_j + z_{j+1} = b_j,\ j = 2, 3, \cdots, q-1, \\ z_{q-1} + Az_q = b_q, \end{cases} \quad (4.5.33)$$

from the three consecutive equations

$$\begin{array}{lllll} z_{j-2} + Az_{j-1} + z_j & & & & = b_{j-1}, \\ z_{j-1} + Az_j + z_{j+1} & & & & = b_j, \\ z_j + Az_{j+1} + z_{j+2} & & & & = b_{j+1}. \end{array}$$

One can for all even $j = 2, 4, \ldots$ eliminate z_{j-1} and z_{j+1} by subtracting A times the second equation from the sum of the others:

$$z_{j-2} + (2I - A^2)z_j + z_{j+2} = b_{j-1} - Ab_j + b_{j+1}.$$

For q odd, we obtain the reduced system

$$\begin{bmatrix} 2I - A^2 & I & & 0 \\ I & 2I - A^2 & \ddots & \\ & \ddots & \ddots & I \\ 0 & & I & 2I - A^2 \end{bmatrix} \begin{bmatrix} z_2 \\ z_4 \\ \vdots \\ z_{q-1} \end{bmatrix} = \begin{bmatrix} b_1 + b_3 - Ab_2 \\ b_3 + b_5 - Ab_4 \\ \vdots \\ b_{q-2} + b_q - Ab_{q-1} \end{bmatrix}. \quad (4.5.34)$$

A solution $\{z_2, z_4, ..., z_{q-1}\}$ of (4.5.34) is known, then $\{z_1, z_3, \ldots\}$ can be determined by (from(4.5.33)):

$$\begin{bmatrix} A & & & 0 \\ & A & & \\ & & \ddots & \\ 0 & & & A \end{bmatrix} \begin{bmatrix} z_1 \\ z_3 \\ \vdots \\ z_q \end{bmatrix} = \begin{bmatrix} b_1 - z_2 \\ b_3 - z_2 - z_4 \\ \vdots \\ b_q - z_{q-1} \end{bmatrix}. \tag{4.5.35}$$

Thus, (4.5.34) has the same structure as (4.5.32):

$$M^{(1)} z^{(1)} = b^{(1)}$$

with

$$M^{(1)} = \begin{bmatrix} A^{(1)} & I & & 0 \\ I & A^{(1)} & \ddots & \\ & \ddots & \ddots & I \\ 0 & & I & A^{(1)} \end{bmatrix}, \quad A^{(1)} \equiv 2I - A^2,$$

$$z^{(1)} = \begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ \vdots \\ z_{q_1}^{(1)} \end{bmatrix} \equiv \begin{bmatrix} z_2 \\ z_4 \\ \vdots \\ z_{q-1} \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_{q_1}^{(1)} \end{bmatrix} \equiv \begin{bmatrix} b_1 + b_3 - Ab_2 \\ b_3 + b_5 - Ab_4 \\ \vdots \\ b_{q-2} + b_q - Ab_{q-1} \end{bmatrix},$$

so that the reduction procedure just described can be applied to $M^{(1)}$ again. In general, for $q = q_0 = 2^{k+1} - 1$, we obtain a sequence of $A^{(r)}$ and $b^{(r)}$ according to:

$$\begin{cases} \text{Set } A^{(0)} = A, b^{(0)} = b_j,\ j = 1, 2, ..., q_0,\ q^{(0)} = q = 2^{k+1} - 1. \\ \\ \text{For } r = 0, 1, 2, \ldots, k-1: \\ \\ \quad (1)\ A^{(r+1)} \equiv 2I - (A^{(r)})^2, \\ \\ \quad (2)\ b_j^{(r+1)} \equiv b_{2j-1}^{(r)} + b_{2j+1}^{(r)} - A^{(r)} b_{2j}^{(r)},\ j = 1, 2, \ldots, 2^{k-r} - 1\,(\equiv q_{r+1}). \end{cases} \tag{4.5.36}$$

For each stage $r+1$, $r = 0, 1, ..., k-1$, one has a linear system

$$M^{(r+1)} z^{(r+1)} = b^{(r+1)}$$

or

$$\begin{bmatrix} A^{(r+1)} & I & & 0 \\ I & A^{(r+1)} & \ddots & \\ & \ddots & \ddots & I \\ 0 & & I & A^{(r+1)} \end{bmatrix} \begin{bmatrix} z_1^{(r+1)} \\ z_2^{(r+1)} \\ \vdots \\ z_{q_{r+1}}^{(r+1)} \end{bmatrix} = \begin{bmatrix} b_1^{(r+1)} \\ b_2^{(r+1)} \\ \vdots \\ b_{q_{r+1}}^{(r+1)} \end{bmatrix}.$$

Its solution $z^{(r+1)}$ furnishes the subvectors with even indices of $z^{(r)}$ of the system $M^{(r)}z^{(r)} = b^{(r)}$ in stage r,

$$\begin{bmatrix} z_2^{(r)} \\ z_4^{(r)} \\ \vdots \\ z_{q_{r-1}}^{(r)} \end{bmatrix} \equiv \begin{bmatrix} z_1^{(r+1)} \\ z_2^{(r+1)} \\ \vdots \\ z_{q_{r+1}}^{(r+1)} \end{bmatrix},$$

while the subvector with odd indices of $z^{(r)}$ can be obtained by solving

$$\begin{bmatrix} A^{(r)} & & & \\ & A^{(r)} & & \\ & & \ddots & \\ & & & A^{(r)} \end{bmatrix} \begin{bmatrix} z_1^{(r)} \\ z_3^{(r)} \\ \vdots \\ z_{q_r}^{(r)} \end{bmatrix} = \begin{bmatrix} b_1^{(r)} - z_2^{(r)} \\ b_3^{(r)} - z_2^{(r)} - z_4^{(r)} \\ \vdots \\ b_{q_r}^{(r)} - z_{q_{r-1}}^{(r)} \end{bmatrix}.$$

From $A^{(r)}$, $b^{(r)}$ produced by (4.5.36), the solution $z := z^{(0)}$ of (4.5.32) is thus obtained by the following procedure (13.37) (say!):

Algorithm 4.5.1 **(0)** *Initialization: Determine* $z^{(k)} = z_1^{(k)}$ *by solving* $A^{(k)}z^{(k)} = b^{(k)} = b_1^{(k)}$.

(1) *For* $r = k-1, k-2, ..., 0,$

(a) *Put* $z_{2j}^{(r)} := z_j^{(r+1)},\ j = 1, 2, \ldots, q_{r+1} = 2^{k-r} - 1,$

(b) *For* $j = 1, 3, 5, \ldots, q_r,$ *compute* $z_j^{(r)}$ *by solving*

$$A^{(r)}z_j^{(r)} = b_j^{(r)} - z_{j-1}^{(r)} - z_{j+1}^{(r)} \quad (z_0^{(r)} := z_{q_r+1}^{(r)} := 0).$$

(2) *Put* $z := z^{(0)}$.

Remark 4.5.1 *(4.5.36) and Algorithm 4.5.1 are still unsatisfactory, as it has serious numerical drawbacks. We have the following disadvantages:*

(1) $A^{(r+1)} = 2I - (A^{(r)})^2$ *in (I) of (4.5.36) is very expensive, the tridiagonal matrix* $A^{(0)} = A$ *as* r *increases, very quickly turns into a dense matrix. So that, the computation of* $(A^{(r)})^2$ *and the solution of (1b) of Algorithm 4.5.1 become very expensive.*

(2) *The magnitude of* $A^{(r)}$ *grows exponentially: For*

$$A = A^0 = \begin{bmatrix} -4 & 1 & & 0 \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -4 \end{bmatrix}, \quad \| A^0 \| \geq 4, \| A^{(r)} \| \approx \| A^{(r-1)} \|^2 \geq 4^{2^r}.$$

Both drawbacks can be avoided by a suitable reformulation of the algorithm. The explicit computation of $A^{(r)}$ *is avoided if one exploits the fact that* $A^{(r)}$ *can be represented as a product of tridiagonal matrices.*

Theorem 4.5.3 *One has for all $r \geq 0$,*

$$A^{(r)} = -\prod_{j=1}^{2^r}[-(A + 2cos\theta_j^{(r)} \cdot I)],$$

where $\theta_j^{(r)} := (2j-1)\pi/2^{r+1}$, for $j = 1, 2, \ldots, 2^r$.

Proof: By (1) of (4.5.36), one has $A^{(0)} = A$, $A^{(r+1)} = 2I - (A^{(r)})^2$, so that there exists a polynomial $P_r(t)$ of degree 2^r such that

$$A^{(r)} = P_r(A). \tag{4.5.37}$$

Evidently, P_r satisfy

$$\begin{aligned} P_0(t) &= t, \\ P_{r+1}(t) &= 2 - (P_r(t))^2, \end{aligned}$$

so that P_r has the form

$$P_r(t) = -(-t)^{2^r} + \cdots . \tag{4.5.38}$$

By induction, using the substitution $t = -2\cos\theta$, we get

$$P_r(-2\cos\theta) = -2\cos(2^r\theta). \tag{4.5.39}$$

The formula is trivial for $r = 0$. If it is valid for some $r \geq 0$, then it is also valid for $r+1$, since

$$\begin{aligned} P_{r+1}(-2\cos\theta) &= 2 - (P_r(-2\cos\theta))^2 \\ &= 2 - 4\cos^2(2^r\theta) \\ &= -2\cos(2 \cdot 2^r\theta). \end{aligned}$$

In view of (4.5.39), $P_r(t)$ has the 2^r distinct real zeros

$$t_j = -2\cos(\frac{2j-1}{2^{r+1}}\pi), \quad j = 1, 2, \ldots, 2^r,$$

and therefore by (4.5.38), the product representation

$$P_r(t) = -\prod_{j=1}^{2^r}[-(t - t_j)].$$

From this, by virtue of (4.5.37), the assertion of Theorem follows immediately. ■

In practice, to reduce the systems $A^{(r)}u = b$ in (1b) of Algorithm 4.5.1 with $A^{(r)}$, recursively to the solution of 2^r systems with tridiagonal matrices

$$A_j^{(r)} := -A - 2\cos\theta_j^{(r)} \cdot I, \quad j = 1, 2, \ldots, 2^r,$$

as follows:

$$\left\{ \begin{array}{ll} A_1^{(r)}u_1 = b & \Rightarrow u_1 \\ A_2^{(r)}u_2 = u_1 & \Rightarrow u_2 \\ \vdots & \\ A_{2^r}^{(r)}u_{2^r} = u_{2^r-1} & \Rightarrow u_{2^r} \quad \Rightarrow u := -u_{2^r}. \end{array} \right. \tag{4.5.40}$$

Remark 4.5.2 **(i)** *It is easily verified, the tridiagonal matrices $A_j^{(r)}$ are positive definite. One can use Cholesky decomposition for the systems.*

(ii) *The numerical instability which occurs in (4.5.36)(2) because of the exponential growth of $A^{(r)}$ can be avoided.*

Buneman (1969) suggested that by introducing in place of the $b_j^{(r)}$ other vectors $p_j^{(r)}, q_j^{(r)}, j = 1, 2, ..., q_r$, which are related to $b_j^{(r)}$:

$$b_j^{(r)} = A^{(r)} p_j^{(r)} + q_j^{(r)}, \quad j = 1, 2, ..., q_r, \tag{4.5.41}$$

which can be computed as follows:

$$\left\{ \begin{array}{l} \text{Set } p_j^{(0)} := 0, q_j^{(0)} := b_j = b_j^{(0)}, j = 1, 2, ..., q_r. \\ \\ \text{For } r = 0, 1, .., k-1 : \\ \\ \quad \text{for } j = 1, 2, ..., q_{r+1} : \text{ Compute} \\ \\ \quad (1)\ p_j^{(r+1)} := p_{2j}^{(r)} - (A^{(r)})^{-1}[p_{2j-1}^{(r)} + p_{2j+1}^{(r)} + q_{2j}^{(r)}], \\ \\ \quad (2)\ q_j^{(r+1)} := q_{2j-1}^{(r)} + q_{2j+1}^{(r)} - 2p_j^{(r+1)}. \end{array} \right. \tag{4.5.42}$$

The computation of $p_j^{(r+1)}$ in (4.5.42)(1) is as in (4.5.40). The solution u of $A^{(r)}u = p_{2j-1}^{(r)} + p_{2j+1}^{(r)} - q_{2j}^{(r)}$ with the factorization of $A^{(r)}$ in Theorem 4.5.3 and then computing $p_j^{(r+1)}$ from u by means of

$$p_j^{(r+1)} := p_{2j}^{(r)} - u.$$

Let us prove by induction on r that $p_j^{(r)}$, $q_j^{(r)}$ in (4.5.42) satisfy the relation (4.5.41). For $r = 0$ (4.5.41) is trivial. Assume that (4.5.41) holds true for some $r \geq 0$. Because of (4.5.36)(2) and $A^{(r+1)} = 2I - (A^{(r)})^2$, we then have

$$\begin{aligned} b_j^{(r+1)} &= b_{2j+1}^{(r)} + b_{2j-1}^{(r)} - A^{(r)} b_{2j}^{(r)} \\ &= A^{(r)} p_{2j+1}^{(r)} + q_{2j+1}^{(r)} + A^{(r)} p_{2j-1}^{(r)} + q_{2j-1}^{(r)} - A^{(r)}[A^{(r)} p_{2j}^{(r)} + q_{2j}^{(r)}] \\ &= A^{(r)}[p_{2j+1}^{(r)} + p_{2j-1}^{(r)} - q_{2j}^{(r)}] + A^{(r+1)} p_{2j}^{(r)} + q_{2j-1}^{(r)} + q_{2j+1}^{(r)} - 2p_{2j}^{(r)} \\ &= A^{(r+1)} p_{2j}^{(r)} + (A^{(r)})^{-1}\{[2I - A^{(r+1)}][p_{2j+1}^{(r)} + p_{2j-1}^{(r)} - q_{2j}^{(r)}]\} + q_{2j-1}^{(r)} + q_{2j+1}^{(r)} - 2p_{2j}^{(r)} \\ &= A^{(r+1)}\{p_{2j}^{(r)} - (A^{(r)})^{-1}[p_{2j+1}^{(r)} + p_{2j-1}^{(r)} - q_{2j}^{(r)}]\} + q_{2j-1}^{(r)} + q_{2j+1}^{(r)} - 2p_j^{(r+1)} \\ &= A^{(r+1)} p_j^{(r+1)} + q_j^{(r+1)}. \end{aligned}$$

By (4.5.41) we can express $b_j^{(r)}$ in Algorithm 4.5.1 in terms of $p_j^{(r)}, q_j^{(r)}$ and obtain, for example, from (1b) of Algorithm 4.5.1 for $z_j^{(r)}$ the system

$$A^{(r)} z_j^{(r)} = A^{(r)} p_j^{(r)} + q_j^{(r)} - z_{j-1}^{(r)} - z_{j+1}^{(r)},$$

which can be solved by determining u of

$$A^{(r)}u = q_j^{(r)} - z_{j-1}^{(r)} - z_{j+1}^{(r)},$$

and put $z_j^{(r)} := u + p_j^{(r)}$. Replacing the $b_j^{(r)}$ in (4.5.36) and Algorithm 4.5.1 systematically by $p_j^{(r)}$ and $q_j^{(r)}$ one obtains:

Algorithm 4.5.2 (Algorithm of Buneman) *Consider the system (4.5.32), with* $q = 2^{k+1} - 1$.
(0) *Initialization: Put* $p_j^{(0)} := 0$, $q_j^{(0)} := b_j$, $j = 1, 2, \ldots, q_0 := q$.
(1) *For* $r = 0, 1, \ldots, k-1$,

For $j = 1, 2, ...,\quad q_{r+1} := 2^{k-r} - 1$:

Compute u *of* $A^{(r)}u = p_{2j-1}^{(r)} + p_{2j+1}^{(r)} - q_{2j}^{(r)}$ *by the factorization of Theorem 4.5.3 and put* $p_j^{(r+1)} := p_{2j}^{(r)} - u$, $q_j^{(r)} := q_{2j-1}^{(r)} + q_{2j+1}^{(r)} - 2p_j^{(r+1)}$.

(2) *Determine* u *of the systems* $A^{(k)}u = q_1^{(k)}$, *and put* $z^{(k)} := z_1^{(k)} := p_1^{(k)} + u$.
(3) *For* $r = k-1, k-2, ..., 0$,

(a) Put $z_{2j}^{(r)} := z_j^{(r+1)}$ *for* $j = 1, 2, ..., q_{r+1}$.
(b) For $j = 1, 3, 5, ..., q_r$ *determine the solution* u *of* $A^{(r)}u = q_j^{(r)} - z_{j-1}^{(r)} - z_{j+1}^{(r)}$ *and put* $z_j^{(r)} := p_j^{(r)} + u$.

(4) *Put* $z := z^{(0)}$.

Remark 4.5.3 *This method is very efficient: For the model problem (4.1.14)* ($a = 1 = b$, $p = q = N = 2^{k+1} - 1$), *with its* N^2 *unknowns, on requires about* $3kN^2 \approx 3N^2 \log_2 N$ *multiplications and about the same number of additions.*

4.5.3 Comparison with Iterative Methods

Consider the special model problem

$$\begin{cases} -u_{xx} - u_{yy} = 2\pi^2 \sin \pi x \sin \pi y, \text{ for } (x, y) \in \Omega, \\ u(x, y) = 0, \text{ for } (x, y) \in \partial\Omega, \end{cases}, \tag{4.5.43}$$

where $\Omega = \{(x, y) | 0 < x, y < 1\}$, which has the exact solution $u(x, y) = \sin \pi x \sin \pi y$. Using the discretization we have

$$\begin{cases} Az = b, \ A \text{ as in (4.1.18)}, \\ b = 2h^2\pi^2 \hat{u} \end{cases} \tag{4.5.44}$$

with $\hat{u} := [\hat{u}_{11}, \hat{u}_{21}, \ldots, \hat{u}_{N1}, \ldots, \hat{u}_{1N}, \ldots, \hat{u}_{NN}]^T$ and $\hat{u}_{ij} := u(x_i, y_j) = \sin i\pi h \sin j\pi h$, $h = 1/(N+1)$.

Method	k	N	$r^{(i)}$	i
Jacobi		5	3.5×10^{-3}	60
		10	1.2×10^{-3}	235
Gauss-Seidel		5	3.0×10^{-3}	33
		10	1.1×10^{-3}	127
		25	5.6×10^{-3}	600
Relaxation		5	1.6×10^{-3}	13
		10	0.9×10^{-3}	28
		25	0.6×10^{-3}	77
		50	1.0×10^{-2}	180
ADI	2	5	0.7×10^{-3}	9
		10	4.4×10^{-3}	12
		25	2.0×10^{-2}	16
	4	5	1.2×10^{-3}	9
		10	0.8×10^{-3}	13
		25	1.6×10^{-5}	14
		50	3.6×10^{-4}	14

Table 4.1: Comparison results for Jacobi, Gauss-Seidel, SOR and ADI methods

Remark 4.5.4 *The vector b in (4.5.44) is an eigenvector of $J = (4I - A)/4$, also an eigenvector of A. We have $Jb = \mu b$ with $\mu = \cos \pi h$. The exact solution of (4.5.44) can be found*

$$z := \frac{h^2\pi^2}{2(1 - \cos \pi h)}\hat{u}. \tag{4.5.45}$$

As a measure for the error we took the residual, weighted by $1/h^2$:

$$\hat{r}^{(r)} := \frac{1}{h^2} \parallel Az^{(i)} - b \parallel_\infty .$$

We start with $z^{(0)} := 0 \quad (\hat{r}^{(0)} = 2\pi^2 \approx 20)$. We show the table computed by Jacobi, Gauss-Seidel, SOR and ADI methods respectively:

Since the Algorithm of Buneman in §13.2 is a noniterative method which yields the exact solution of (4.5.44) in a finite number of steps at the expense of about $3N^2 \log_2 N$ multiplications. From (4.5.45), by Taylor expansion in powers of h, we have

$$z - \hat{u} = \left(\frac{\pi^2 h^2}{2(1-\cos \pi h)} - 1 \right) \hat{u} = \frac{h^2\pi^2}{12}\hat{u} + O(h^4),$$

so that the error $\|z - \hat{u}\|_\infty$, in as much as $\|\hat{u}\|_\infty \leq 1$, satisfies $\|z - \hat{u}\|_\infty \leq \frac{h^2\pi^2}{12} + O(h^4)$. In order to compute z with an error of the order h^2, the needed number of iterations and operations for the Jacobi, Gauss-Seidel and SOR methods are shown in Table 4.2.
For a given N, $R^{(m)}_{ADI}$ is minimized for $m \approx \ln[4(N+1)/\pi]$, in which case $\sqrt[m]{4(N+1)/\pi} \approx e$. The ADI method with optimal choice of m and optimal choice of parameter thus requires

$$R^{(m)}_{ADI} \log_{10}(N+1)^2 \approx 3.60(\log_{10} N)^2$$

Method	No. of iterations	No. of operations
Jacobi	$0.467(N+1)^2\log_{10}(N+1)^2 \approx N^2\log_{10}N$	$5N^4\log_{10}N$
Gauss-Seidel	$0.234(N+1)^2\log_{10}(N+1)^2 \approx 0.5N^2\log_{10}N$	$2.5N^4\log_{10}N$
Optimal SOR	$0.36(N+1)\log_{10}(N+1)^2 \approx 0.72N\log_{10}N$	$3.6N^3\log_{10}N$

Table 4.2: Number of iterations and operations for Jacobi, Gauss-Seidel and SOR methods

iterations to approximate the solution z of (4.5.44) with error h^2. The ADI requires about $8N^2$ multiplications per iteration, so that the total number of operations is about

$$28.8N^2(\log_{10}N)^2.$$

The Buneman method, according to §13.2 requires only $3N^2\log_2 N \approx 10N^2\log_{10}N$ multiplications for the computation of the exact solution of (4.5.44). This clearly shows the superiority of Buneman method.

4.6 Derivation and Properties of the Conjugate Gradient Method

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite (s.p.d.) matrix. Here n is very large and A is sparse. Consider the linear system

$$Ax = b.$$

4.6.1 A Variational Problem, Steepest Descent Method (Gradient Method).

Consider the functional $F : \mathbb{R}^n \to \mathbb{R}$ with

$$F(x) = \frac{1}{2}x^TAx - b^Tx = \frac{1}{2}\sum_{i,k=1}^{n} a_{ik}x_ix_k - \sum_{i=1}^{n} b_ix_i. \tag{4.6.1}$$

Then it holds:

Theorem 4.6.1 *For a vector x^* the following statements are equivalent:*

$$\begin{array}{ll}(i) & F(x^*) < F(x),\ \textit{for all } x \neq x^*,\\ (ii) & Ax^* = b.\end{array} \tag{4.6.2}$$

Proof: From assumption there exists $z_0 = A^{-1}b$ and $F(x)$ can be rewritten as

$$F(x) = \frac{1}{2}(x - z_0)^TA(x - z_0) - \frac{1}{2}z_0^TAz_0. \tag{4.6.3}$$

Since A is positive definite, $F(x)$ has a minimum at $x = z_0$ and only at $x = z_0$, it follows the assertion. ■

Therefore, the solution of the linear system $Ax = b$ is equal to the solution of the minimization problem

$$F(x) = \frac{1}{2}x^TAx - b^Tx = \min!. \tag{4.6.4}$$

Method of the steepest descent

Let x_k be an approximate of the exact solution x^* and p_k be a search direction. We want to find an α_{k+1} such that

$$F(x_k + \alpha_{k+1} p_k) < F(x_k).$$

Set $x_{k+1} := x_k + \alpha_{k+1} p_k$. This leads to the following basic problem.

Basic Problem: Given x, $p \neq 0$, find α_0 such that

$$\Phi(\alpha_0) = F(x + \alpha_0 p) = \min!$$

Solution: Since

$$\begin{aligned} F(x+\alpha p) &= \frac{1}{2}(x+\alpha p)^T A(x+\alpha p) - b^T(x+\alpha p) \\ &= \frac{1}{2}\alpha^2 p^T A p + \alpha(p^T A x - p^T b) + F(x), \end{aligned}$$

it follows that if we take

$$\alpha_0 = \frac{(b-Ax)^T p}{p^T A p} = \frac{r^T p}{p^T A p}, \tag{4.6.5}$$

where $r = b - Ax = -\mathrm{grad}F(x) =$ residual, then $x + \alpha_0 p$ is the minimal solution. Moreover,

$$F(x + \alpha_0 p) = F(x) - \frac{1}{2}\frac{(r^T p)^2}{p^T A p}. \tag{4.6.6}$$

Steepest Descent Method with Optimal Choice α_{k+1} (Determine α_k via the given data x_0, p_0, $p_1, \cdots$): Let

$$x_{k+1} = x_k + \frac{r_k^T p_k}{p_k^T A p_k} p_k,\ r_k = b - Ax_k, \tag{4.6.7}$$

$$F(x_{k+1}) = F(x_k) - \frac{1}{2}\frac{(r_k^T p_k)^2}{p_k^T A p_k},\ k = 0, 1, 2, \cdots. \tag{4.6.8}$$

Then, it holds

$$r_{k+1}^T p_k = 0. \tag{4.6.9}$$

Since

$$\frac{d}{d\alpha} F(x_k + \alpha p_k) = \mathrm{grad}F(x_k + \alpha p_k)^T p_k,$$

as in (4.6.5) $\alpha_{k+1} = \frac{r_k^T p_k}{p_k^T A p_k}$, it follows that $\mathrm{grad}F(x_k + \alpha_{k+1} p_k)^T p_k = 0$. Thus

$$(b - Ax_{k+1})^T p_k = r_{k+1}^T p_k = 0,$$

hence (4.6.9) holds.

Steepest Descent Method (Gradient Method)

Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a differential function on x. Then it holds

$$\frac{\Phi(x + \varepsilon p) - \Phi(x)}{\varepsilon} = \Phi'(x)^T p + O(\varepsilon).$$

The right hand side takes minimum at $p = -\frac{\Phi'(x)}{\|\Phi'(x)\|}$ (i.e., the largest descent) for all p with $\|p\| = 1$ (neglect $O(\varepsilon)$). Hence, it suggests to choose

$$p_k = -\text{grad}F(x_k) = b - Ax_k. \tag{4.6.10}$$

Gradient Method:

$$\begin{cases} \text{Given } x_0, \text{ for } k = 1, 2, \cdots \\ r_{k-1} = b - Ax_{k-1}, \text{ if } r_{k-1} = 0, \text{ then stop; else} \\ \alpha_k = \frac{r_{k-1}^T r_{k-1}}{r_{k-1}^T A r_{k-1}}, \ x_k = x_{k-1} + \alpha_k r_{k-1}. \end{cases} \tag{4.6.11}$$

Cost in each step: compute Ax_{k-1} (Ar_{k-1} does not need to compute).

To prove the convergence of Gradient method, we need the Kontorowitsch inequality: Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$. Then it holds

$$\sum_{i=1}^{n} \alpha_i \lambda_i \sum_{j=1}^{n} \alpha_j \lambda_j^{-1} \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1 \lambda_n} = \frac{1}{4}(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}})^2. \tag{4.6.12}$$

Proof of (4.6.12): Consider the n points $P_i = (\lambda_i, \ 1/\lambda_i)$. Let B be the region between $y = 1/x$ and the straight line through $P_1, \ P_n$. The slope of the straight line $\overleftrightarrow{P_1 P_n}$ is

$$\frac{1/\lambda_n - 1/\lambda_1}{\lambda_n - \lambda_1} = -\frac{1}{\lambda_n \lambda_1}.$$

The point $P = \sum_{i=1}^{n} \alpha_i P_i$ lies in B. Maximize xy, for all $(x, y) \in B$. The point (ξ, η) which lies on $\overline{P_1 P_n}$ is a maximum for $\xi\eta$ and has the coordinates:

$$\xi = \alpha\lambda_n + (1-\alpha)\lambda_1, \text{ and } \eta = \alpha\frac{1}{\lambda_n} + (1-\alpha)\frac{1}{\lambda_1}.$$

Since

$$\begin{aligned} 0 &= \frac{d}{d\alpha}[(\alpha\lambda_n + (1-\alpha)\lambda_1)(\alpha\frac{1}{\lambda_n} + (1-\alpha)\frac{1}{\lambda_1})] \\ &= \frac{d}{d\alpha}[\alpha^2 + (1-\alpha)^2 + \alpha(1-\alpha)(\frac{\lambda_n}{\lambda_1} + \frac{\lambda_1}{\lambda_n})] \\ &= 2\alpha + 2(\alpha - 1) + (1 - 2\alpha)(\frac{\lambda_n}{\lambda_1} + \frac{\lambda_1}{\lambda_n}) \\ &= (1 - 2\alpha)(\frac{\lambda_n}{\lambda_1} + \frac{\lambda_1}{\lambda_n} - 2), \end{aligned}$$

it follows $\alpha = 1/2$. Hence

$$\xi\eta = \frac{1}{4}(\lambda_1 + \lambda_n)(\frac{1}{\lambda_1} + \frac{1}{\lambda_n}) = \frac{1}{4}\frac{(\lambda_1 + \lambda_n)^2}{\lambda_1 \lambda_n}.$$

So (4.6.12) holds. ■

Another form: Let A be s.p.d. (symmetric positive definite) and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ be the eigenvalues of A. Let x be a vector with $\| x \|_2^2 = x^T x = 1$, then it holds

$$x^T Ax \cdot x^T A^{-1} x \leq \frac{1}{4} \frac{(\lambda_1 + \lambda_n)^2}{\lambda_1 \lambda_n} = \frac{1}{4} (\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}})^2. \tag{4.6.13}$$

Proof of (4.6.13): Let U be an orthogonal matrix satisfying $UAU^T = \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$. Then we have

$$x^T Ax = x^T U^T \Lambda U x = y^T \Lambda y = \sum_{i=1}^{n} y_i^2 \lambda_i \quad (y := Ux).$$

Similarly,

$$x^T A^{-1} x = y^T \Lambda^{-1} y = \sum_{i=1}^{n} y_i^2 \frac{1}{\lambda_i}.$$

From (4.6.12) follows (4.6.13). ■

Theorem 4.6.2 *If x_k, x_{k-1} are two approximations of the gradient method (4.6.11) for solving $Ax = b$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ are the eigenvalues of A, then it holds:*

$$F(x_k) + \frac{1}{2} b^T A^{-1} b \leq (\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n})^2 [F(x_{k-1}) + \frac{1}{2} b^T A^{-1} b], \tag{4.6.14a}$$

i.e.,

$$\| x_k - x^* \|_A \leq (\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}) \| x_{k-1} - x^* \|_A, \tag{4.6.14b}$$

where $\|x\|_A = \sqrt{x^T Ax}$. Thus the gradient method is convergent.

Proof: By computation,

$$\begin{aligned}
F(x_k) \quad + \quad & \frac{1}{2} b^T A^{-1} b = \frac{1}{2} (x_k - x^*)^T A (x_k - x^*) \\
= \quad & \frac{1}{2} (x_{k-1} - x^* + \alpha_k r_{k-1})^T A (x_{k-1} - x^* + \alpha_k r_{k-1}) \quad (\text{ since } A(x_{k-1} - x^*) = -r_{k-1}) \\
= \quad & \frac{1}{2} [(x_{k-1} - x^*)^T A (x_{k-1} - x^*) - 2\alpha_k r_{k-1}^T r_{k-1} + \alpha_k^2 r_{k-1}^T A r_{k-1}] \\
= \quad & \frac{1}{2} [r_{k-1}^T A^{-1} r_{k-1} - \frac{(r_{k-1}^T r_{k-1})^2}{r_{k-1}^T A r_{k-1}}] \\
= \quad & \frac{1}{2} r_{k-1}^T A^{-1} r_{k-1} [1 - \frac{(r_{k-1}^T r_{k-1})^2}{r_{k-1}^T A r_{k-1} \cdot r_{k-1}^T A^{-1} r_{k-1}}] \\
\leq \quad & \frac{1}{2} r_{k-1}^T A^{-1} r_{k-1} [1 - \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}] \quad (\text{ from (4.6.13)}) \\
= \quad & [F(x_{k-1}) + \frac{1}{2} b^T A^{-1} b] (\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n})^2.
\end{aligned}$$

■

If the condition number of A ($= \lambda_1 / \lambda_n$) is large, then $\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \approx 1$. The gradient method converges very slowly. Hence this method is not recommendable.

4.6.2 Conjugate gradient method

It is favorable to choose that the search directions $\{p_i\}$ as mutually A-conjugate, where A is symmetric positive definite.

Definition 4.6.1 *Two vectors p and q are called A-conjugate (A-orthogonal), if $p^TAq = 0$.*

Remark 4.6.1 *Let A be symmetric positive definite. Then there exists a unique s.p.d. B such that $B^2 = A$. Denote $B = A^{1/2}$. Then $p^TAq = (A^{1/2}p)^T(A^{1/2}q)$.*

Lemma 4.6.3 *Let $p_0, \ldots, p_r \neq 0$ be pairwisely A-conjugate. Then they are linearly independent.*

Proof: From $0 = \sum_{j=0}^{r} c_j p_j$ follows that

$$p_k^T A(\sum_{j=0}^{r} c_j p_j) = 0 = \sum_{j=0}^{r} c_j p_k^T A p_j = c_k p_k^T A p_k,$$

so $c_k = 0$, for $k = 1, \ldots, r$. ■

Theorem 4.6.4 *Let A be s.p.d. and $p_0, \ldots, p_{n-1}$ be nonzero pairwisely A-conjugate vectors. Then*

$$A^{-1} = \sum_{j=0}^{n-1} \frac{p_j p_j^T}{p_j^T A p_j}. \tag{4.6.15}$$

Remark 4.6.2 *$A = I$, $U = (p_0, \ldots, p_{n-1})$, $p_i^T p_i = 1$, $p_i^T p_j = 0$, $i \neq j$. $UU^T = I$ and $I = UU^T$. Then*

$$I = (p_0, \ldots, p_{n-1}) \begin{bmatrix} p_0^T \\ \vdots \\ p_{n-1}^T \end{bmatrix} = p_0 p_0^T + \cdots + p_{n-1} p_{n-1}^T.$$

Proof of Theorem 4.6.4: Since $\tilde{p_i} = \frac{A^{1/2}p_i}{\sqrt{p_i^T A p_i}}$ are orthonormal, for $i = 0, 1, \ldots, n-1$, we have

$$\begin{aligned} I &= \tilde{p_0}\tilde{p_0}^T + \ldots + \tilde{p}_{n-1}\tilde{p}_{n-1}^T \\ &= \sum_{i=0}^{n-1} \frac{A^{1/2}p_i p_i^T A^{1/2}}{p_i^T A p_i} = A^{1/2}\left(\sum_{i=0}^{n-1} \frac{p_i p_i^T}{p_i^T A p_i}\right) A^{1/2}. \end{aligned}$$

Thus,

$$A^{-1/2} I A^{-1/2} = A^{-1} = \sum_{i=0}^{n-1} \frac{p_i p_i^T}{p_i^T A p_i}.$$

■

Remark 4.6.3 *Let $Ax^* = b$ and x_0 be an arbitrary vector. Then from $x^* - x_0 = A^{-1}(b - Ax_0)$ and (4.6.15) follows that*

$$x^* = x_0 + \sum_{i=0}^{n-1} \frac{p_i^T(b - Ax_0)}{(p_i^T A p_i)} p_i. \tag{4.6.16}$$

Theorem 4.6.5 *Let A be s.p.d. and $p_0, \ldots, p_{n-1} \in \mathbb{R}^n \backslash \{0\}$ be pairwisely A-orthogonal. Given x_0 and let $r_0 = b - Ax_0$. For $k = 0, \ldots, n-1$, let*

$$\alpha_k = \frac{p_k^T r_k}{p_k^T A p_k}, \tag{4.6.17}$$

$$x_{k+1} = x_k + \alpha_k p_k, \tag{4.6.18}$$

$$r_{k+1} = r_k - \alpha_k A p_k. \tag{4.6.19}$$

Then the following statements hold:

(i) $r_k = b - Ax_k$. (By induction).

(ii) x_{k+1} minimizes $F(x)$ (see (4.6.1)) on $x = x_k + \alpha p_k$, $\alpha \in \mathbb{R}$.

(iii) $x_n = A^{-1}b = x^*$.

(iv) x_k minimizes $F(x)$ on the affine subspace $x_0 + S_k$, where $S_k = Span\{p_0, \ldots, p_{k-1}\}$.

Proof: (i): By Induction and using (4.6.18) (4.6.19).
(ii): From (4.6.5) and (i).
(iii): It is enough to show that x_k (which defined in (4.6.18)) corresponds with the partial sum in (4.6.16), i.e.,

$$x_k = x_0 + \sum_{\nu=0}^{k-1} \frac{p_\nu^T(b - Ax_0)}{p_\nu^T A p_\nu} p_\nu.$$

Then it follows that $x_n = x^*$ from (4.6.16). From (4.6.17) and (4.6.18) we have

$$x_k = x_0 + \sum_{\nu=0}^{k-1} \alpha_\nu p_\nu = x_0 + \sum_{\nu=0}^{k-1} \frac{p_\nu^T(b - Ax_\nu)}{p_\nu^T A p_\nu} p_\nu.$$

To show that

$$p_\nu^T(b - Ax_\nu) = p_\nu^T(b - Ax_0). \tag{4.6.20}$$

From $x_k - x_0 = \sum_{\nu=0}^{k-1} \alpha_\nu p_\nu$ we obtain

$$p_k^T A x_k - p_k^T A x_0 = \sum_{\nu=0}^{k-1} \alpha_\nu p_k^T A p_\nu = 0.$$

So (4.6.20) holds.
(iv): From (4.6.19) and (4.6.17) follows that

$$p_k^T r_{k+1} = p_k^T r_k - \alpha_k p_k^T A p_k = 0.$$

From (4.6.18), (4.6.19) and by the fact that $r_{k+s} - r_{k+s+1} = \alpha_{k+s} A p_{k+s}$ and p_k are orthogonal (for $s \geq 1$) follows that

$$p_k^T r_{k+1} = p_k^T r_{k+2} = \ldots = p_k^T r_n = 0.$$

Hence we have

$$p_i^T r_k = 0,\ i = 0, \ldots, k-1,\ k = 1, 2, \ldots, n. \text{ (i.e., } i < k). \tag{4.6.21}$$

We now consider $F(x)$ on $x_0 + S_k$:

$$F(x_0 + \sum_{i=0}^{k-1} \xi_i p_i) = \varphi(\xi_0, \ldots, \xi_{k-1}).$$

$F(x)$ is minimal on $x_0 + S_k$ if and only if all derivatives $\frac{\partial \varphi}{\partial \xi_i}$ vanish at x. But

$$\frac{\partial \varphi}{\partial \xi_s} = [\text{grad} F(x_0 + \sum_{i=0}^{k-1} \xi_i p_i)]^T p_s,\ s = 0, 1, \ldots, k-1. \tag{4.6.22}$$

If $x = x_k$, then $\text{grad} F(x) = -r_k$. From (4.6.21) follows that

$$\frac{\partial \varphi}{\partial \xi_s}(x_k) = 0,\ \text{for } s = 0, 1, \ldots, k-1.$$

Another proof of (iv): For arbitrary $d \in \mathbb{R}^n$ it holds

$$\begin{aligned} F(x_0 + d) - F(x_0) &= \frac{1}{2}(x_0 + d)^T A(x_0 + d) - b^T(x_0 + d) - \frac{1}{2} x_0^T A x_0 + b^T x_0 \\ &= \frac{1}{2} d^T A d - d^T(b - A x_0). \end{aligned}$$

So for $d = \sum_{i=0}^{k-1} \xi_i p_i$ we have

$$\begin{aligned} F(x_0 + \sum_{i=0}^{k-1} \xi_i p_i) &= F(x_0) + \frac{1}{2}(\sum_{i=0}^{k-1} \xi_i p_i)^T A(\sum_{j=0}^{k-1} \xi_j p_j) - \sum_{i=0}^{k-1} \xi_i p_i^T (b - A x_0) \\ &= F(x_0) + \frac{1}{2} \sum_{i=0}^{k-1} [\xi_i^2 p_i^T A p_i - 2 p_i^T (b - A x_0) \xi_i] = \text{min!}. \end{aligned} \tag{4.6.23}$$

The equation (4.6.23) holds if and only if

$$\xi_i^2 p_i^T A p_i - 2 \xi_i p_i (b - A x_0) = \text{min!}\ i = 0, \ldots, k-1,$$

if and only if

$$\xi_i = \frac{p_i^T (b - A x_0)}{p_i^T A p_i} = \frac{p_i^T r_i}{p_i^T A p_i} = \alpha_i$$

from (4.6.20) and (4.6.17). Thus $x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i p_i$ minimizes F on $x_0 + \text{span}\{p_0, \ldots, p_{k-1}\}$.

■

Remark 4.6.4 *The following conditions are equivalent:* (i) $p_i^T A p_j = 0$, $i \neq j$, *A-conjugate,* (ii) $p_i^T r_k = 0$, $i < k$, (iii) $r_i^T r_j$, $i \neq j$.

Proof of (iii):

$$p_i^T r_k = 0 \quad \Leftrightarrow \quad (r_i^T + \beta_{i-1} p_{i-1}^T) r_k, \ i < k \quad \Leftrightarrow \quad r_i^T r_k = 0, \ i < k \quad \Leftrightarrow \quad r_i^T r_j = 0, \quad i \neq j.$$

Remark 4.6.5 *It holds*

$$< p_0, p_1, \cdots, p_k > = < r_0, r_1, \cdots, r_k > = < r_0, Ar_0, \cdots, A^k r_0 >$$

Since $p_1 = r_1 + \beta_0 p_0 = r_1 + \beta_0 r_0$, $r_1 = r_0 - \alpha_0 A r_0$, *by induction, we have*

$$r_2 = r_1 - \alpha_0 A p_1 = r_1 - \alpha_0 A(r_1 + \beta_0 r_0) = r_0 - \alpha_0 A r_0 - \alpha_0 A(r_0 - \alpha_0 A r_0 + \beta_0 r_0).$$

Algorithm 4.6.1 (Method of conjugate directions) *Let A be s.p.d., b and $x_0 \in \mathbb{R}^n$. Given $p_0, \ldots, p_{n-1} \in \mathbb{R}^n \backslash \{0\}$ pairwisely A-orthogonal.*

$$\begin{aligned}
&r_0 = b - A x_0, \\
&\text{For } k = 0, \ldots, n-1, \\
&\quad \alpha_k = \tfrac{p_k r_k}{p_k^T A p_k}, \ x_{k+1} = x_k + \alpha_k p_k, \\
&\quad r_{k+1} = r_k - \alpha_k A p_k = b - A x_{k+1}, \\
&\text{end for}
\end{aligned}$$

From Theorem 4.6.5 we get $x_n = A^{-1} b$.

4.6.3 Practical Implementation

In the k-th step a direction p_k which is A-orthogonal to $p_0, \ldots, p_{k-1}$ must be determined. It allows for A-orthogonalization of r_k against $p_0, \ldots, p_{k-1}$ (see (4.6.21)). Let $r_k \neq 0$, $F(x)$ decreases strictly in the direction $-r_k$. For $\varepsilon > 0$ small, we have $F(x_k - \varepsilon r_k) < F(x_k)$. It follows that F takes its minimum at a point ($\neq x_k$) on $x_0 +$ span$\{p_0, \ldots, p_{k-1}, r_k\}$. So it holds $x_{k+1} \neq x_k$, i.e., $\alpha_k \neq 0$. This derives that Conjugate Gradient method.

Algorithm 4.6.2 (Conjugate Gradient method (CG-method), (Stiefel-Hestenes, 1952)) *Let A be s.p.d., $b \in \mathbb{R}^n$, choose $x_0 \in \mathbb{R}^n$, $r_0 = b - A x_0 = p_0$.*
If $r_0 = 0$, then $N = 0$ stop; otherwise for $k = 0, 1, \ldots$

$$\begin{aligned}
&(a)\ \alpha_k = \frac{p_k^T r_k}{p_k^T A p_k}, \\
&(b)\ x_{k+1} = x_k + \alpha_k p_k, \\
&(c)\ r_{k+1} = r_k - \alpha_k A p_k = b - A x_{k+1}, \ \text{if } r_{k+1} = 0, \ \text{let } N = k+1, \ \text{stop.} \qquad (4.6.24) \\
&(d)\ \beta_k = \frac{-r_{k+1}^T A p_k}{p_k^T A p_k}, \\
&(e)\ p_{k+1} = r_{k+1} + \beta_k p_k.
\end{aligned}$$

Theorem 4.6.6 *The CG-method holds*

(i) *If k steps of CG-method are executable, i.e., $r_i \neq 0$, for $i = 0, \ldots, k$, then $p_i \neq 0$, $i \leq k$ and $p_i^T A p_j = 0$ for $i, j \leq k$, $i \neq j$.*

(ii) *The CG-method breaks down after N steps for $r_N = 0$ and $N \leq n$.*

(iii) $x_N = A^{-1}b$.

Proof: (i): By induction on k, it is trivial for $k = 0$. Suppose that (i) is true until k and $r_{k+1} \neq 0$. Then p_{k+1} is well-defined. we want to verify that (a) $p_{k+1} \neq 0$, (b) $p_{k+1}^T A p_j = 0$, for $j = 0, 1, \ldots, k$.
For (a): First, it holds $r_{k+1}^T p_k = r_k^T p_k - \alpha_k p_k^T A p_k = 0$ by (4.6.24)(c). Let $p_{k+1} = 0$. Then from (4.6.24)(e) we have $r_{k+1} = -\beta_k p_k \neq 0$. So, $\beta_k \neq 0$, hence $0 = r_{k+1}^T p_k = -\beta_k p_k^T p_k \neq 0$. This is a contradiction, so $p_{k+1} \neq 0$.
For (b): From (4.6.24)(d) and (e), we have

$$p_{k+1}^T A p_k = r_{k+1}^T A p_k + \beta_k p_k^T A p_k = 0.$$

Let $j < k$, from (4.6.24)(e) we have

$$p_{k+1}^T A p_j = r_{k+1}^T A p_j + \beta_k p_k^T A p_j = r_{k+1}^T A p_j. \tag{4.6.25}$$

It is enough to show that

$$A p_j \in \text{span}\{p_0, ..., p_{j+1}\},\ j < k. \tag{4.6.26}$$

Then from the relation $p_i^T r_j = 0$, $i < j \leq k+1$, which has been proved in (4.6.21), follows (b).
Claim (4.6.26): For $r_j \neq 0$, it holds that $\alpha_j \neq 0$. (4.6.24)(c) shows that

$$A p_j = \frac{1}{\alpha_j}(r_j - r_{j+1}) \in \text{span}\{r_0, ..., r_{j+1}\}.$$

(4.6.24)(e) shows that $\text{span}\{r_0, ..., r_{j+1}\} = \text{span}\{p_0, ..., p_{j+1}\}$ with $r_0 = p_0$, so is (4.6.26).
(ii): Since $\{p_i\}_{i=0}^{k+1} \neq 0$ and are mutually A-orthogonal, $p_0, ..., p_{k+1}$ are linearly independent. Hence there exists a $N \leq n$ with $r_N = 0$. This follows $x_N = A^{-1}b$. ■
Advantage:(1) Break-down in finite steps. (2) Less cost in each step: one matrix $\times$ vector.

4.6.4 Convergence of CG-method

Consider the following A-norm with A being s.p.d.

$$\|x\|_A = (x^T A x)^{1/2}. \tag{4.6.27}$$

Let $x^* = A^{-1}b$. Then from (4.6.3) we have

$$F(x) - F(x^*) = \frac{1}{2}(x - x^*)^T A (x - x^*) = \frac{1}{2}\|x - x^*\|_A^2, \tag{4.6.28}$$

where x_k is the k-th iterate of CG-method. From Theorem 4.6.5 x_k minimizes the functional F on x_0+ span$\{p_0, ..., p_{k-1}\}$. Hence it holds

$$\|x_k - x^*\|_A \leq \|y - x^*\|_A,\ y \in x_0 + \text{span}\{p_0, .., p_{k-1}\}. \tag{4.6.29}$$

From (4.6.24)(c)(e) it is easily seen that both p_k and r_k can be written as linear combination of r_0, $Ar_0, \ldots, A^{k-1}r_0$. If $y \in x_0 + \text{span}\{p_0, \ldots, p_{k-1}\}$, then

$$y = x_0 + c_1 r_0 + c_2 A r_0 + ... + c_k A^{k-1} r_0 = x_0 + \mathcal{P}_{k-1}(A) r_0,$$

where $\mathcal{P}_{k-1}$ is a polynomial of degree $\leq k-1$. But $r_0 = b - Ax_0 = A(x^* - x_0)$, thus

$$\begin{aligned} y - x^* &= (x - x^*) + \mathcal{P}_{k-1}(A)A(x^* - x_0) \\ &= [I - A\mathcal{P}_{k-1}(A)]\,(x_0 - x^*) = \tilde{\mathcal{P}}_k(A)(x_0 - x^*), \end{aligned} \tag{4.6.30}$$

where degree $\tilde{\mathcal{P}}_k \leq k$ and

$$\tilde{\mathcal{P}}_k(0) = 1. \tag{4.6.31}$$

Conversely, if $\tilde{\mathcal{P}}_k$ is a polynomial of degree $\leq k$ and satisfies (4.6.31), then

$$x^* + \tilde{\mathcal{P}}_k(A)(x_0 - x^*) \in x_0 + S_k.$$

Hence (4.6.29) means that if $\tilde{\mathcal{P}}_k$ is a polynomial of degree $\leq k$ with $\tilde{\mathcal{P}}_k(0) = 1$, then

$$\|x_k - x^*\|_A \leq \|\tilde{\mathcal{P}}_k(A)(x_0 - x^*)\|_A. \tag{4.6.32}$$

Lemma 4.6.7 *Let A be s.p.d. It holds for every polynominal Q_k of degree k that*

$$\max_{x \neq 0} \frac{\|Q_k(A)x\|_A}{\|x\|_A} = \rho(Q_k(A)) = \max\{|Q_k(\lambda)| : \lambda \textit{ eigenvalue of } A\}. \tag{4.6.33}$$

Proof:

$$\begin{aligned} \frac{\|Q_k(A)x\|_A^2}{\|x\|_A^2} &= \frac{x^T Q_k(A) A Q_k(A) x}{x^T A x} \\ &= \frac{(A^{1/2}x)^T Q_k(A) Q_k(A)(A^{1/2}x)}{(A^{1/2}x)(A^{1/2}x)} \qquad (\text{Let } z := A^{1/2}x) \\ &= \frac{z^T Q_k(A)^2 z}{z^T z} \leq \rho(Q_k(A)^2) = \rho^2(Q_k(A)). \end{aligned}$$

Equality holds for suitable x, hence the first equality is shown. The second equality holds by the fact that $Q_k(\lambda)$ is an eigenvalue of $Q_k(A)$, where λ is an eigenvalue of A. ■

From (4.6.33) we have that

$$\|x_k - x^*\|_A \leq \rho(\tilde{\mathcal{P}}_k(A))\|x_0 - x^*\|_A, \tag{4.6.34}$$

where degree $\tilde{\mathcal{P}}_k \leq k$ and $\tilde{\mathcal{P}}_k(0) = 1$.

Replacement problem for (4.6.34): For $0 < a < b$,

$$\text{minmax}\{|\mathcal{P}_k(\lambda)| : a \leq \lambda \leq b,\ \text{for all polynomials of degree} \leq k \text{ with } \mathcal{P}_k(0) = 1\} \tag{4.6.35}$$

(if $a = 0$, it is clearly $\min\max\{|\ \mathcal{P}_k(\lambda)\ |\} = 1$). We use Chebychev polynomials of the first kind for the solution. They are defined by

$$\begin{cases} T_0(t) = 1, T_1(t) = t, \\ T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t). \end{cases} \tag{4.6.36}$$

it holds $T_k(\cos\phi) = \cos(k\phi)$ by using $\cos((k+1)\phi) + \cos((k-1)\phi) = 2\cos\phi\cos k\phi$. Especially,

$$T_k(\cos\frac{j\pi}{k}) = \cos(j\pi) = (-1)^j, \text{ for } j = 0, \ldots, k,$$

i.e. T_k takes maximal value "one" at $k+1$ positions in $[-1, 1]$ with alternating sign. In addition (Exercise!), we have

$$T_k(t) = \frac{1}{2}[(t + \sqrt{t^2-1})^k + (t - \sqrt{t^2-1})^k]. \tag{4.6.37}$$

Lemma 4.6.8 *The solution of the problem (4.6.35) is given by*

$$Q_k(t) = T_k\left(\frac{2t-a-b}{b-a}\right) \Big/ T_k\left(\frac{a+b}{a-b}\right),$$

i.e., for all $\mathcal{P}_k$ of degree $\leq k$ with $\mathcal{P}_k(0) = 1$ it holds

$$\max_{\lambda\in[a,b]} |Q_k(\lambda)| \leq \max_{\lambda\in[a,b]} |\mathcal{P}_k(\lambda)|.$$

Proof: $Q_k(0) = 1$. If t runs through the interval $[a, b]$, then $(2t - a - b)/(b - a)$ runs through the interval $[-1, 1]$. Hence, in $[a, b]$, $Q_k(t)$ has $k + 1$ extreme with alternating sign and absolute value $\delta = |T_k(\frac{a+b}{a-b})^{-1}|$.

If there are a $\mathcal{P}_k$ with $\max\{|\mathcal{P}_k(\lambda)| : \lambda \in [a, b]\} < \delta$, then $Q_k - \mathcal{P}_k$ has the same sign as Q_k of the extremal values, so $Q_k - \mathcal{P}_k$ changes sign at $k + 1$ positions. Hence $Q_k - \mathcal{P}_k$ has k roots, in addition a root zero. This contradicts that degree$(Q_k - \mathcal{P}_k) \leq k$. ■

Lemma 4.6.9 *It holds*

$$\delta = \left|T_k\left(\frac{b+a}{a-b}\right)^{-1}\right| = \frac{1}{T_k\left(\frac{b+a}{b-a}\right)} = \frac{2c^k}{1+c^{2k}} \leq 2c^k, \tag{4.6.38}$$

where $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ and $\kappa = b/a$.

Proof: For $t = \frac{b+a}{b-a} = \frac{\kappa+1}{\kappa-1}$, we compute

$$t + \sqrt{t^2-1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} = c^{-1}$$

and

$$t - \sqrt{t^2-1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = c.$$

Hence from (4.6.37) follows

$$\delta = \frac{2}{c^k + c^{-k}} = \frac{2c^k}{1+c^{2k}} \leq 2c^k.$$

■

Theorem 4.6.10 *CG-method satisfies the following error estimate*

$$\|x_k - x^*\|_A \leq 2c^k \|x_0 - x^*\|_A, \tag{4.6.39}$$

where $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, $\kappa = \frac{\lambda_1}{\lambda_n}$ *and* $\lambda_1 \geq \cdots \geq \lambda_n > 0$ *are the eigenvalues of* A.

Proof: From (4.6.34) we have

$$\begin{aligned} \|x_k - x^*\|_A &\leq \rho\left(\mathcal{P}_k(A)\right) \|x_0 - x^*\|_A \\ &\leq \max\left\{|\mathcal{P}_k(\lambda)| : \lambda_1 \geq \lambda \geq \lambda_n\right\} \|x_0 - x^*\|_A, \end{aligned}$$

for all $\mathcal{P}_k$ of degree $\leq k$ with $\mathcal{P}_k(0) = 1$. From Lemma 4.6.8 and Lemma 4.6.9 follows that

$$\begin{aligned} \|x_k - x^*\|_A &\leq \max\left\{|Q_k(\lambda)| : \lambda_1 \geq \lambda \geq \lambda_n\right\} \|x_0 - x^*\|_A \\ &\leq 2c^k \|x_0 - x^*\|_A. \end{aligned}$$

■

Remark 4.6.6 *To compare with Gradient method (see (4.6.14b)): Let* x_k^G *be the* k*th iterate of Gradient method. Then*

$$\|x_k^G - x^*\|_A \leq \left|\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right|^k \|x_0 - x^*\|_A.$$

But

$$\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{\kappa - 1}{\kappa + 1} > \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = c,$$

because in general $\sqrt{\kappa} \ll \kappa$. *Therefore the CG-method is much better than Gradient method.*

4.7 CG-method as an iterative method, preconditioning

Consider the linear system of a symmetric positive definite matrix A

$$Ax = b. \tag{4.7.1}$$

Let C be a nonsingular symmetric matrix and consider a new linear system

$$\tilde{A}\tilde{x} = \tilde{b} \tag{4.7.2}$$

with $\tilde{A} = C^{-1}AC^{-1}$ s.p.d., $\tilde{b} = C^{-1}b$ and $\tilde{x} = Cx$.
Applying CG-method to (4.7.2) it yields:
Choose $\tilde{x}_0$, $\tilde{r}_0 = \tilde{b} - \tilde{A}\tilde{x}_0 = \tilde{p}_0$.
If $\tilde{r}_0 = 0$, stop, otherwise for $k = 0,\ 1,\ 2,\ \ldots$,

$$\left\{\begin{array}{ll} (a) & \tilde{\alpha}_k = \tilde{p}_k^T \tilde{r}_k / \tilde{p}_k^T C^{-1}AC^{-1}\tilde{p}_k, \\ (b) & \tilde{x}_{k+1} = \tilde{x}_k + \tilde{\alpha}_k \tilde{p}_k, \\ (c) & \tilde{r}_{k+1} = \tilde{r}_k - \tilde{\alpha}_k C^{-1}AC^{-1}\tilde{p}_k, \\ & \text{if } \tilde{r}_{k+1} = 0 \text{ stop; otherwise,} \\ (d) & \tilde{\beta}_k = -\tilde{r}_{k+1}^T C^{-1}AC^{-1}\tilde{p}_k / \tilde{p}_k C^{-1}AC^{-1}\tilde{p}_k, \\ (e) & \tilde{p}_{k+1} = \tilde{r}_{k+1} + \tilde{\beta}_k \tilde{p}_k. \end{array}\right. \tag{4.7.3}$$

Simplification: Let

$$C^{-1}\tilde{p}_k = p_k, \quad x_k = C^{-1}\tilde{x}_k, \quad z_k = C^{-1}\tilde{r}_k.$$

Then

$$r_k = C\tilde{r}_k = C\left(\tilde{b} - \tilde{A}\tilde{x}_k\right) = C\left(C^{-1}b - C^{-1}AC^{-1}Cx_k\right) = b - Ax_k.$$

Algorithm 4.7.1 (Preconditioned CG-method (PCG))
$M = C^2$, *choose* $x_0 = C^{-1}\tilde{x}_0$, $r_0 = b - Ax_0$, *solve* $Mp_0 = r_0$.
If $r_0 = 0$ *stop, otherwise for* $k = 0,\ 1,\ 2, ...,$

$$\left\{\begin{array}{ll} (a) & \alpha_k = p_k^T r_k / p_k^T A p_k, \\ (b) & x_{k+1} = x_k + \alpha_k p_k, \\ (c) & r_{k+1} = r_k - \alpha_k A p_k, \\ & \text{if } r_{k+1} = 0, \text{ stop; otherwise } Mz_{k+1} = r_{k+1}, \\ (d) & \beta_k = -z_{k+1}^T A p_k / p_k^T A p_k, \\ (e) & p_{k+1} = z_{k+1} + \beta_k p_k. \end{array}\right. \tag{4.7.4}$$

Algorithm 4.7.1 is CG-method with preconditioner M. If $M = I$, then it is CG-method.
Additional cost per step: solve one linear system $Mz = r$ for z.
Advantage: $\mathrm{cond}(M^{-1/2}AM^{-1/2}) \ll \mathrm{cond}(A)$.

4.7.1 A new point of view of PCG

From (4.6.21) and Theorem 4.6.6 follows that ${p_i}^T r_k = 0$ for $i < k$, i.e., $({r_i}^T + \beta_{i-1}{p_{i-1}}^T)r_k = {r_i}^T r_k = 0$, $i < k$ and ${p_i}^T A p_j = 0$, $i \neq j$. That is, the CG method requires ${r_i}^T r_j = 0,\ i \neq j$. So, the PCG method satisfies ${p_i}^T C^{-1}AC^{-1}p_j = 0 \Leftrightarrow \tilde{r}_j^T \tilde{r}_j = 0, \quad i \neq j$ and requires

$$\begin{aligned} z_i^T M z_j &= r_i^T M^{-1} M M^{-1} r_j = {r_i}^T M^{-1} r_j \\ &= \left({r_i}^T C^{-1}\right)\left(C^{-1} r_j\right) = \widetilde{r_i}^T \widetilde{r_j} = 0, \quad i \neq j. \end{aligned}$$

Consider the iteration (in two parameters):

$$x_{k+1} = x_{k-1} + \omega_{k+1}\left(\alpha_k z_k + x_k - x_{k-1}\right) \tag{4.7.5}$$

with α_k and ω_{k+1} being two undetermined parameters. Let $A = M - N$. Then from $Mz_k = r_k \equiv b - Ax_k$ follows that

$$\begin{aligned} Mz_{k+1} &= b - A\left(x_{k-1} + \omega_{k+1}\left(\alpha_k z_k + x_k - x_{k-1}\right)\right) \\ &= Mz_{k-1} - \omega_{k+1}\left[\alpha_k (M - N) z_k + M(z_{k-1} - z_k)\right] \end{aligned} \tag{4.7.6}$$

For PCG method $\{\alpha_k, \omega_{k+1}\}$ are computed so that

$${z_p}^T M z_q = 0, \quad p \neq q,\ p, q = 0, 1, \cdots, n-1. \tag{4.7.7}$$

Since $M > 0$, there is some $k \leq n$ such that $z_k = 0$. Thus, $x_k = x$, the iteration converges no more than n steps. We show that (4.7.7) holds by induction. Assume

$${z_p}^T M z_q = 0, \quad p \neq q,\ p, q = 0, 1, \cdots, k \tag{4.7.8}$$

holds until k. If we choose

$$\alpha_k = {z_k}^T M z_k \Big/ {z_k}^T (M-N) z_k,$$

then

$${z_k}^T M z_{k+1} = 0$$

and if we choose

$$\omega_{k+1} = \left(1 - \alpha_k \frac{z_{k-1}^T N z_k}{z_{k-1}^T M z_{k-1}}\right)^{-1}, \tag{4.7.9}$$

then

$$z_{k-1}^T M z_{k+1} = 0.$$

We want to simplify ω_{k+1}. From (4.7.6) follows that

$$M z_k = M z_{k-2} - \omega_k \left(\alpha_{k-1}(M-N) z_{k-1} + M(z_{k-2} - z_{k-1})\right). \tag{4.7.10}$$

Multiplying (4.7.10) by ${z_k}^T$ and from (4.7.8) we get

$${z_k}^T N z_{k-1} = {z_k}^T M z_k \Big/ \omega_k \alpha_{k-1}. \tag{4.7.11}$$

Since $z_{k-1}^T N z_k = {z_k}^T Z z_{k-1}$, from (4.7.11) the equation (4.7.9) becomes

$$\omega_{k+1} = \left(1 - \frac{\alpha_k {z_k}^T M z_k}{\alpha_{k-1} z_{k-1}^T M z_{k-1}} \frac{1}{\omega_k}\right)^{-1}. \tag{4.7.12}$$

From (4.7.6) for $j < k-1$ we have

$${z_j}^T M z_{k+1} = \alpha_k \omega_{k+1} {z_j}^T N z_k. \tag{4.7.13}$$

But (4.7.6) holds for $j < k-1$,

$$M z_{j+1} = M z_{j-1} - \omega_{j+1} \left(\alpha_j (M-N) z_j + M(z_{j-1} - z_j)\right). \tag{4.7.14}$$

Multiplying (4.7.14) by ${z_k}^T$ we get

$${z_k}^T N z_j = 0.$$

Since $N = N^T$, it follows that

$${z_j}^T M z_{k+1} = 0, \quad \text{for} \quad j < k-1.$$

Thus, we proved that ${z_p}^T M z_q = 0$, $p \neq q$, $p,\ q = 0,\ 1,\ \cdots,\ n-1$. ■

Consider (4.7.5) again

$$x_{k+1} = x_{k-1} + \omega_{k+1}(\alpha_k z_k + x_k - x_{k-1}).$$

Since $M z_k = r_k = b - A x_k$, if we set $\omega_{k+1} = \alpha_k = 1$, then

$$x_{k+1} = M^{-1}(b - A x_k) + x_k \equiv x_k + z_k. \tag{4.7.15}$$

Here z_k is referred to as a correction term. Write $A = M - N$. Then (4.7.15) becomes

$$\begin{aligned} Mx_{k+1} &= b - Ax_k + Mx_k \\ &= Nx_k + b. \end{aligned} \tag{4.7.16}$$

Recall the Iterative Improvement in Subsection 2.3.6:

Solve $Ax = b$,
$r_k = b - Ax_k$,
$Az_k = r_k, \leftrightarrow Mz_k = r_k$.
$x_{k+1} = x_k + z_k$.

(i) **Jacobi method** ($\omega_{k+1} = \alpha_k = 1$)**:** $A = D - (L + R)$,

$$x_{k+1} = x_k + D^{-1}(b - Ax_k).$$

(ii) **Gauss-Seidel** ($\omega_{k+1} = \alpha_k = 1$)**:** $A = (D - L) - R$,

$$x_{k+1} = x_k + (D - L)^{-1}(b - Ax_k).$$

i.e.,

$$\begin{aligned} x_j^{(k+1)} &= b_j - \sum_{p=1}^{j-1} a_{jp}x_p^{(k+1)} - \sum_{p=j+1}^{n} a_{jp}x_p^{(k)} + x_j^{(k)} - 1 \cdot x_j^{(k)} \\ &= x_j^{(k)} + b_j - (a_{j1}, \ldots, a_{j,j-1}, 1, a_{j,j+1}, \ldots, a_{jn}) \begin{bmatrix} x_1^{(k+1)} \\ \vdots \\ x_{j-1}^{(k+1)} \\ x_j^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix}, \quad (D = I). \end{aligned}$$

(iii) **SOR-method** ($\omega_{k+1} = 1, \alpha_k = \omega$)**:** Solve $\omega Ax = \omega b$. Write

$$\omega A = (D - \omega L) - ((1 - \omega)D + \omega R) \equiv M - N.$$

Then

$$\begin{aligned} x_{k+1} &= (D - \omega L)^{-1}(\omega R + (1 - \omega)D)x_k + (D - \omega L)^{-1}\omega b \\ &= (D - \omega L)^{-1}((D - \omega L) - \omega A)x_k + (D - \omega L)^{-1}\omega b \\ &= (I - (D - \omega L)^{-1}\omega A)x_k + (D - \omega L)^{-1}\omega b \\ &= x_k + (D - \omega L)^{-1}\omega(b - Ax_k) \\ &= x_k + \omega M^{-1}r_k. \end{aligned}$$

i.e.,

$$\begin{aligned} x_j^{(k+1)} &= \omega\left(b_j - \sum_{p=1}^{j-1} a_{jp}x_p^{(k+1)} - \sum_{p=j+1}^{n} a_{jp}x_p^{(k)}\right) + (1-\omega)x_j^{(k)} \\ &= x_j^{(k)} + \omega b_j - \omega\left(a_{j1}, \cdots, a_{j,j-1}, 1, a_{j,j+1}, \cdots, a_{jn}\right) \begin{bmatrix} x_1^{(k+1)} \\ \vdots \\ x_{j-1}^{(k+1)} \\ x_j^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix}. \end{aligned}$$

(iv) **Chebychev Semi-iterative method (later!)** $(\omega_{k+1} = \omega_{k+1}, \alpha_k = \gamma)$**:**

$$x_{k+1} = x_{k-1} + \omega_{k+1}\left(\gamma z_k + x_k - x_{k-1}\right).$$

We can think of the scalars ω_{k+1}, α_k in (4.7.5) as acceleration parameters that can be chosen to speed the convergence of the iteration $Mx_{k+1} = Nx_k + b$. Hence any iterative method based on the splitting $A = M - N$ can be accelerated by the Conjugate Gradient Algorithm so long as M (the preconditioner) is symmetric and positive definite.

Choices of M (Criterion):

(i) $\text{cond}(M^{-1/2}AM^{-1/2})$ is nearly by 1, i.e., $M^{-1/2}AM^{-1/2} \approx I, A \approx M$.

(ii) The linear system $Mz = r$ must be easily solved. e.g. $M = LL^T$ (see Section 16.)

(iii) M is symmetric positive definite.

Explanation: Why we need to use preconditioning for solving the linear system $Ax = b$.
Fixed Point Principle:

$$\begin{aligned} x &= b - Ax + x \\ &= (I - A)x + b. \end{aligned}$$

Thus $x = Bx + b$ with $B \equiv I - A$.
Fixed Point Iteration:

$$x_{i+1} = Bx_i + b.$$

Let $e_i = x_i - x$. Then $e_{i+1} = Be_i = B^i e_0$. Thus $\{e_i\} \to 0$ if and only if $\rho(B) < 1$. Hence we want to find an M so that $M^{-1}A \approx I$ with $A = M - N$. Consider

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

then

$$\begin{aligned} x_{i+1} &= \left(I - M^{-1}A\right)x_i + M^{-1}b \\ &= \left(I - M^{-1}(M - N)\right)x_i + M^{-1}b, \\ &= M^{-1}Nx_i + M^{-1}b. \end{aligned} \tag{4.7.17}$$

Here $A = M - N$ is called a splitting iterative scheme and $Mz = r$ should be easily solvable. The iteration (4.7.17) is called a preconditioned fixed point iteration.

Jacobi: $A = D - (L + R)$.

Gauss-Seidel: $A = (D - L) - R$.

SOR (Successive Over Relaxation): $Ax = b$, $\omega Ax = \omega b$, $(\omega > 1)$,

$$\begin{aligned} \omega A &= \omega D - \omega L - \omega R \\ &= (D - \omega L) - [(1-\omega)D + \omega R] \\ &\equiv M - N. \end{aligned}$$

This implies,

$$\begin{aligned} x_{i+1} &= (D - \omega L)^{-1}[(1-\omega)D + \omega R]x_i + (D - \omega L)^{-1}\omega b \\ &= M^{-1}Nx_i + M^{-1}\omega b \\ &\quad (M^{-1}N = I - (D - \omega L)^{-1}\omega A). \end{aligned}$$

SSOR (Symmetric Successive Over Relaxation): A is symmetric and $A = D - L - L^T$. Let

$$\begin{cases} M_\omega : = D - \omega L, \\ N_\omega : = (1-\omega)D + \omega L^T, \end{cases} \quad \text{and} \quad \begin{cases} M_\omega^T = D - \omega L^T, \\ N_\omega^T = (1-\omega)D + \omega L. \end{cases}$$

Then from the iterations

$$\begin{aligned} M_\omega x_{i+1/2} &= N_\omega x_i + \omega b, \\ M_\omega^T x_{i+1} &= N_\omega^T x_{i+1/2} + \omega b, \end{aligned}$$

follows that

$$\begin{aligned} x_{i+1} &= \left(M_\omega^{-T} N_\omega^T M_\omega^{-1} N_\omega\right) x_i + \tilde{b} \\ &\equiv Gx_i + \omega \left(M_\omega^{-T} N_\omega^T M_\omega^{-1} + M_\omega^{-T}\right) b \\ &\equiv Gx_i + M(\omega)^{-1} b. \end{aligned}$$

But

$$\begin{aligned} &((1-\omega)D + \omega L)(D - \omega L)^{-1} + I \\ &= (\omega L - D - \omega D + 2D)(D - \omega L)^{-1} + I \\ &= -I + (2-\omega)D(D - \omega L)^{-1} + I \\ &= (2-\omega)D(D - \omega L)^{-1}, \end{aligned}$$

Thus

$$M(\omega)^{-1} = \omega \left(D - \omega L^T\right)^{-1} (2-\omega)D(D - \omega L)^{-1},$$

then

$$\begin{aligned} M(\omega) &= \frac{1}{\omega(2-\omega)}(D - \omega L)D^{-1}\left(D - \omega L^T\right) \\ &\approx (D - L)D^{-1}\left(D - L^T\right), \ (\omega = 1). \end{aligned} \tag{4.7.18}$$

For a suitable ω the condition number of $M(\omega)^{-1/2}AM(\omega)^{-1/2}$, i.e., $\mathrm{cond}(M(\omega)^{-1/2}AM(\omega)^{-1/2})$, can be considered smaller than $\mathrm{cond}(A)$. Axelsson(1976) showed (without proof): Let

$$\mu = \max_{x\neq 0} \frac{x^T Dx}{x^T Ax} \ (\leq \mathrm{cond}(A))$$

and

$$\delta = \max_{x\neq 0} \frac{x^T (LD^{-1}L^T - \frac{1}{4}D)x}{x^T Ax} \geq \frac{1}{4}.$$

Then

$$\mathrm{cond}\left(M(\omega)^{-1/2}AM(\omega)^{-1/2}\right) \leq \frac{1 + \frac{(2-\omega)^2}{4\omega} + \omega\delta}{2\omega} = \kappa(\omega)$$

for $\omega^* = \frac{2}{1+2\sqrt{(2\delta+1)/2\mu}}$, $\kappa(\omega^*)$ is minimal and $\kappa(\omega^*) = 1/2 + \sqrt{(1/2+\delta)\mu}$. Especially

$$\mathrm{cond}\left(M(\omega^*)^{-1/2}AM(\omega^*)^{-1/2}\right) \leq \frac{1}{2} + \sqrt{(1/2+\delta)\mathrm{cond}(A)} \sim \sqrt{\mathrm{cond}(A)}.$$

Disadvantage : μ, δ in general are unknown.

SSOR + Conjugate Gradient method.
SSOR + Chebychev Semi-iterative Acceleration (later!)

4.8 Incomplete Cholesky Decomposition

Let A be sparse and symmetric positive definite. Consider the Cholesky decomposition of $A = LL^T$. L is a lower triangular matrix with $l_{ii} > 0$ $(i = 1, ..., n)$. L can be heavily occupied (fill-in). Consider the following decomposition

$$A = LL^T - N, \tag{4.8.1}$$

where L is a lower triangular matrix with prescribed reserved pattern E and N is "small".

Reserved Pattern: $E \subset \{1, ..., n\} \times \{1, ..., n\}$ with $\begin{cases} (i,i) \in E, \ i = 1, ..., n \\ (i,j) \in E \ \Rightarrow \ (j,i) \in E \end{cases}$

For a given reserved pattern E we construct the matrices L and N as in (4.8.1) with

$$\text{(i)} \quad A = LL^T - N, \tag{4.8.2a}$$

$$\text{(ii)} \quad L: \ \text{lower triangular with } l_{ii} > 0 \text{ and } l_{ij} \neq 0 \ \Rightarrow \ (i,j) \in E, \tag{4.8.2b}$$

$$\text{(iii)} \quad N = (n_{ij}), \ n_{ij} = 0, \ \text{if } (i,j) \in E \tag{4.8.2c}$$

First step: Consider the Cholesky decomposition of A,

$$A = \begin{pmatrix} a_{11} & a_1^T \\ a_1 & A_1 \end{pmatrix} = \begin{pmatrix} \sqrt{a_{11}} & 0 \\ a_1/\sqrt{a_{11}} & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \overline{A}_1 \end{pmatrix} \begin{pmatrix} \sqrt{a_{11}} & a_1^T/\sqrt{a_{11}} \\ 0 & I \end{pmatrix},$$

where $\overline{A}_1 = A_1 - a_1 a_1^T/a_{11}$. Then

$$A = L_1 \begin{pmatrix} 1 & 0 \\ 0 & \overline{A}_1 \end{pmatrix} L_1^T.$$

For the Incomplete Cholesky decomposition the first step will be so modified. Define $b_1 = (b_{21}, ..., b_{n1})^T$ and $c_1 = (c_{21}, ..., c_{n1})^T$ by

$$b_{j1} = \begin{cases} a_{j1}, & (j,1) \in E, \\ 0, & \text{otherwise}, \end{cases} \qquad c_{j1} = b_{j1} - a_{j1} = \begin{cases} 0, & (j,1) \in E, \\ -a_{j1}, & \text{otherwise}. \end{cases} \tag{4.8.3}$$

Then

$$A = \begin{pmatrix} a_{11} & b_1^T \\ b_1 & A_1 \end{pmatrix} - \begin{pmatrix} 0 & c_1^T \\ c_1 & 0 \end{pmatrix} = \tilde{B}_0 - C_1. \tag{4.8.4}$$

Compute the Cholesky decomposition on $\tilde{B}$, we get

$$\tilde{B}_0 = \begin{pmatrix} \sqrt{a_{11}} & 0 \\ b_1/\sqrt{a_{11}} & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \bar{B}_1 \end{pmatrix} \begin{pmatrix} \sqrt{a_{11}} & b_1^T/\sqrt{a_{11}} \\ 0 & I \end{pmatrix} = L_1 B_1 L_1^T \tag{4.8.5}$$

and

$$\bar{B}_1 = A_1 - \frac{b_1 {b_1}^T}{a_{11}}. \tag{4.8.6}$$

Then

$$A = L_1 B_1 {L_1}^T - C_1. \tag{4.8.7}$$

Consequently, compute the Cholesky decomposition on B_1:

$$B_1 = L_2 B_2 L_2^T - C_2$$

with

$$L_2 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & * & \cdots & \cdots & 0 \\ \vdots & * & 1 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & * & \cdots & \cdots & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 0 & & \cdots & \cdots & 0 \\ & 0 & * & \cdots & * \\ \vdots & * & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & * & \cdots & \cdots & 0 \end{pmatrix}.$$

Thus,

$$A = L_1 L_2 B_2 L_2^T L_1^T - L_1 C_2 L_1^T - C_1 \tag{4.8.8}$$

and so on, hence

$$A = L_1 \cdots L_n I L_n^T \cdots L_1^T - C_{n-1} - C_{n-2} - \cdots - C_1 \tag{4.8.9}$$

with

$$L = L_1 \cdots L_n \text{ and } N = C_1 + C_2 + \cdots + C_n. \tag{4.8.10}$$

Lemma 4.8.1 *Let A be s.p.d. E be a reserved patten. Then there is at most a decomposition $A = LL^T - N$, which satisfies the conditions:*

(4.8.2b) : *L is lower triangular with $l_{ii} > 0$, $l_{ii} \neq 0 \Longrightarrow (i,j) \in E$.*
(4.8.2c) : *$N = (n_{ij})$, $n_{ij} = 0$, if $(i,j) \in E$.*

Proof: Let $A = LL^T - N = \bar{L}\bar{L}^T - \bar{N}$. Then $a_{11} = l_{11}^2 = \bar{l}_{11}^2 \Longrightarrow l_{11} = \bar{l_{11}}$ (since l_{11} is positive). Also, $a_{k1} = l_{k1}l_{11} - n_{k1} = \bar{l}_{k1}l_{11} - \bar{n}_{k1}$, so we have

$$\text{If } (k,1) \in E \Longrightarrow n_{k1} = \bar{n}_{k1} = 0 \Longrightarrow l_{k1} = \bar{l}_{k1} = a_{k1}/l_{11}, \tag{4.8.11a}$$

$$\text{If } (k,1) \notin E \Longrightarrow l_{k1} = \bar{l}_{k1} = 0 \Longrightarrow n_{k1} = \bar{n}_{k1} = -a_{k1}. \tag{4.8.11b}$$

Suppose that $l_{ki} = \bar{l}_{ki}$, $n_{ki} = \bar{n}_{ki}$, for $k = i, \cdots, n$, $1 \le i \le m-1$. Then from

$$a_{mm} = l_{mm}^2 + \sum_{k=0}^{m-1} l_{mk}^2 = \bar{l}_{mm}^2 + \sum_{k=1}^{m-1} \bar{l}_{mk}^2$$

follows that $l_{mm} = \bar{l}_{mm}$. Also from

$$a_{rm} = l_{rm}l_{mm} + \sum_{k=1}^{m-1} l_{rk}l_{mk} - n_{rm} = \bar{l}_{rm}\bar{l}_{mm} + \sum_{k=0}^{m-1} \bar{l}_{rk}\bar{l}_{mk} - \bar{n}_{rm}$$

and (4.8.11) follows that $n_{rm} = \bar{n}_{rm}$ and $l_{rm} = \bar{l}_{rm}$ $(r \ge m)$. ■

The Incomplete Cholesky decomposition may not exist, if

$$s_m := a_{mm} - \sum_{k=1}^{m-1} (l_{mk})^2 \le 0.$$

Example 4.8.1 *Let*

$$A = \begin{bmatrix} 1 & -1 & 0 & 2 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -3 \\ 2 & 0 & -3 & 10 \end{bmatrix}.$$

The Cholesky decomposition of A follows $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 2 & 2 & -1 & 1 \end{bmatrix}$. *Consider the Incomplete Cholesky decomposition with patten*

$$E = E(A) = \begin{bmatrix} \times & \times & 0 & \times \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \\ \times & 0 & \times & \times \end{bmatrix}.$$

Above procedures (4.8.3)-(4.8.10) can be performed on A until the computation of l_{44} *(see proof of Lemma 4.8.1),*

$$l_{44}^2 = a_{44} - l_{41}^2 - l_{42}^2 - l_{43}^2 = 10 - 9 - 4 = -3.$$

The Incomplete Cholesky decomposition does not exit for this pattern E. Now take

$$E = \begin{pmatrix} \times & \times & 0 & 0 \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix} \Longrightarrow L \text{ exists and } L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -3 & 1 \end{pmatrix}.$$

Find the certain classes of matrices, which have no breakdown by Incomplete Cholesky decomposition. The classes are

M-matrices, H-matrices.

Definition 4.8.1 *$A \in \mathbb{R}^{n\times n}$ is an M-matrix. If there is a decomposition $A = \sigma I - B$ with $B \geq 0$ ($B \geq 0 \Leftrightarrow b_{ij} \geq 0$ for $i, j = 1, ..., n$) and $\rho(B) = \max\{|\lambda| : \lambda$ is an eigenvalue of $B\}$ $< \sigma$.* Equivalence: *$a_{ij} \leq 0$ for $i \neq j$ and $A^{-1} \geq 0$.*

Lemma 4.8.2 *A is symmetric, $a_{ij} \leq 0$, $i \neq j$. Then the following statements are equivalent*

(i) *A is an M-matrix.*

(ii) *A is s.p.d.*

Proof: (i) $\Rightarrow$ (ii): $A = \sigma I - B$, $\rho(B) < \sigma$. The eigenvalues of A have the form $\sigma - \lambda$, where λ is an eigenvalue of B and $|\lambda| < \sigma$. Since λ is real, so $\sigma - \lambda > 0$ for all eigenvalues λ, it follows that A has only positive eigenvalues. Thus (ii) holds.
(ii) $\Rightarrow$ (i): For $a_{ij} \leq 0$, $(i \neq j)$, there is a decomposition $A = \sigma I - B$, $B \geq 0$ (for example $\sigma = \max(a_{ii})$). Claim $\rho(B) < \sigma$. By Perron-Frobenius Theorem 4.1.7, we have that $\rho(B)$ is an eigenvalue of B. Thus $\sigma - \rho(B)$ is an eigenvalue of A, so $\sigma - \rho(B) > 0$. Then (i) holds. ■

Theorem 4.8.3 *Let A be a symmetric M-matrix. Then the Incomplete Cholesky method described in* (4.8.3)-(4.8.10) *is executable and yields a decomposition $A = LL^T - N$, which satisfies* (4.8.2).

Proof: It is sufficient to show that the matrix B_1 constructed by (4.8.3)-(4.8.7) is a symmetric M-matrix.
(i): We first claim: $\tilde{B}_0$ is an M-matrix. $A = \tilde{B}_0 - C_1 \leq \tilde{B}_0$, (since only negative elements are neglected). There is a $k > 0$ such that $A = kI - \hat{A}$, $\tilde{B}_0 = kI - \hat{B}_0$ with $\hat{A} \geq 0$, $\hat{B}_0 \geq 0$, then $\hat{B}_0 \leq \hat{A}$. By Perron-Frobenius Theorem 4.1.7 follows $\rho(\hat{B}_0) \leq \rho(\hat{A}) < k$. This implies that $\tilde{B}_0$ is an M-matrix.
(ii): Thus $\tilde{B}_0$ is positive definite, hence $B_1 = L_1^{-1}\tilde{B}_0\big(L_1^{-1}\big)^T$ is also positive definite. B_1 has nonpositive off-diagonal element, since $\bar{B}_1 = \bar{A}_1 - \frac{b_1 b_1{}^T}{a_{11}}$. Then B_1 is an M-matrix (by Lemma 4.8.2) ■

Definition 4.8.2 *$A \in \mathbb{R}^{n\times n}$. Decomposition $A = B - C$ is called regular, if $B^{-1} \geq 0$, $C \geq 0$ (regular splitting).*

Theorem 4.8.4 *Let $A^{-1} \geq 0$ and $A = B - C$ is a regular decomposition. Then $\rho(B^{-1}C) < 1$. i.e., the iterative method $Bx_{k+1} = Cx_k + b$ for $Ax = b$ is convergent for all x_0.*

Proof: Since $T = B^{-1}C \geq 0$, $B^{-1}(B - C) = B^{-1}A = I - T$, it follows that

$$(I - T)A^{-1} = B^{-1}.$$

Then

$$0 \leq \sum_{i=0}^{k} T^i B^{-1} = \sum_{i=0}^{k} T^i (I-T) A^{-1} = (I - T^{k+1}) A^{-1} \leq A^{-1}.$$

That is, the monotone sequence $\sum_{i=0}^{k} T^i B^{-1}$ is uniformly bounded. Hence $T^k B^{-1} \to 0$ for $k \to \infty$, then $T^k \to 0$ and $\rho(T) < 1$. ■

Theorem 4.8.5 *If $A^{-1} \geq 0$ and $A = B_1 - C_1 = B_2 - C_2$ are two regular decompositions with $0 \leq C_1 \leq C_2$, then it holds $\rho({B_1}^{-1} C_1) \leq \rho({B_2}^{-1} C_2)$.*

Proof: Let $A = B - C$, $A^{-1} \geq 0$. Then

$$\begin{aligned} \rho(B^{-1}C) &= \rho((A+C)^{-1}C) = \rho([A(I + A^{-1}C)]^{-1}C) \\ &= \rho((I + A^{-1}C)^{-1} A^{-1} C) = \frac{\rho(A^{-1}C)}{1 + \rho(A^{-1}C)}. \end{aligned}$$

$$[\lambda \to \frac{\lambda}{1+\lambda} \text{ monotone for } \lambda \geq 0].$$

Because $0 \leq C_1 \leq C_2$ it follows $\rho(A^{-1}C_1) \leq \rho(A^{-1}C_2)$. Then

$$\rho({B_1}^{-1} C_1) = \frac{\rho(A^{-1}C_1)}{1 + \rho(A^{-1}C_1)} \leq \frac{\rho(A^{-1}C_2)}{1 + \rho(A^{-1}C_2)} = \rho({B_2}^{-1} C_2),$$

since $\lambda \to \frac{\lambda}{1+\lambda}$ is monotone for $\lambda > 0$. ■

Theorem 4.8.6 *If A is a symmetric M-matrix, then the decomposition $A = LL^T - N$ according to Theorem 4.8.3 is a regular decomposition.*

Proof: Because each $L_j^{-1} \geq 0$, it follows $(LL^T)^{-1} \geq 0$, (from $(I - le^T)^{-1} = (I + le^T)$, $l \geq 0$). $N = C_1 + C_2 + \cdots + C_{n-1}$ and all $C_i \geq 0$. ■

Definition 4.8.3 *$A \in \mathbb{R}^{n \times n}$ is called an H-matrix, if the matrix $H = H(A)$ which is defined by*

$$h_{ij} = \begin{cases} a_{ii}, & \text{if } i = j, \\ -|a_{ij}|, & \text{if } i \neq j, \end{cases}$$

is an M-matrix.

Theorem 4.8.7 *(Manteuffel) For any symmetric H-matrix A and any symmetric reserved pattern E there exists an uniquely determined Incomplete Cholesky decomposition of A which satisfies (16.2). [Exercise !].*

History:
(i) CG-method, Hestenes-Stiefel (1952).
(ii) CG-method as iterative method, Reid (1971).
(iii) CG-method with preconditioning, Concus-Golub-Oleary (1976).
(iv) Incomplete Cholesky decomposition, Meijerink-Van der Vorst (1977).
(v) Nonsymmetric matrix, H-matrix, Incomplete Cholesky decomposition, Manteufel (1979).
Other preconditioning:

(i) A blockform $A = [A_{ij}]$ with A_{ij} blocks. Take $M = \text{diag}[A_{11}, ..., A_{kk}]$.

(ii) Try Incomplete Cholesky decomposition: Breakdown can be avoided by two ways. If $z_i = a_{ii} - \Sigma_{k=1}^{i-1} l_{ik}^2 \leq 0$, breakdown, then either set $l_{ii} = 1$ and go on or set $l_{ik} = 0$, $(k = 1, .., i-1)$ until $z_i > 0$ (change reserved pattern E).

(iii) A is an arbitrary nonsingular matrix with all principle determinants $\neq 0$. Then $A = LDR$ exists, where D is diagonal, L and R^T are unit lower triangular. Consider the following generalization of Incomplete Cholesky decomposition.

Theorem 4.8.8 *(Generalization) Let A be an $n \times n$ matrix and E be an arbitrary reserved pattern with $(i,i) \in E$, $i = 1, 2, \ldots, n$. A decomposition of the form $A = LDR - N$ which satisfies:*

(i) L is lower triangular, $l_{ii} = 1$, $l_{ij} \neq 0$, then $(i,j) \in E$,
(ii) R is upper triangular, $r_{ii} = 1$, $r_{ij} \neq 0$, then $(i,j) \in E$,
(iii) D is diagonal $\neq 0$,
(iv) $N = (n_{ij})$, $n_{ij} = 0$ for $(i,j) \in E$.

is uniquely determined. (The decomposition almost exists for all matrices).

4.9 Chebychev Semi-Iteration Acceleration Method

Consider the linear system $Ax = b$. The splitting $A = M - N$ leads to the form

$$x = Tx + f, \; T = M^{-1}N \text{ and } f = M^{-1}b. \tag{4.9.1}$$

The basic iterative method of (4.9.1) is

$$x_{k+1} = Tx_k + f. \tag{4.9.2}$$

How to modify the convergence rate?

Definition 4.9.1 *The iterative method (4.9.2) is called symmetrizable, if there is a matrix W with $detW \neq 0$ and such that $W(I-T)W^{-1}$ is symmetric positive definite.*

Example 4.9.1 *Let A and M be s.p.d., $A = M - N$ and $T = M^{-1}N$, then*

$$I - T = I - M^{-1}N = M^{-1}(M - N) = M^{-1}A.$$

Set $W = M^{1/2}$. Thus,

$$W(I-T)W^{-1} = M^{1/2}M^{-1}AM^{-1/2} = M^{-1/2}AM^{-1/2} \; s.p.d.$$

(i): $M = \text{diag}(a_{ii})$ Jacobi method.
(ii): $M = \frac{1}{\omega(2-\omega)}(D - \omega L)D^{-1}(D - \omega L^T)$ SSOR-method.
(iii): $M = LL^T$ Incomplete Cholesky decomposition.
(iv): $M = I \Rightarrow x_{k+1} = (I - A)x_k + b$ Richardson method.

Lemma 4.9.1 *If (4.9.2) is symmetrizable, then the eigenvalues μ_i of T are real and satisfy*

$$\mu_i < 1, \; \text{for } i = 1, 2, \ldots, n. \tag{4.9.3}$$

Proof: Since $W(I-T)W^{-1}$ is s.p.d., the eigenvalues $1-\mu_i$ of $I-T$ are large than zero. Thus μ_i are real and (4.9.3) holds. ■

Definition 4.9.2 *Let $x_{k+1} = Tx_k + f$ be symmetrizable. The iterative method*

$$\left\{\begin{array}{rcl} u_0 &=& x_0, \\ u_{k+1} &=& \alpha(Tu_k + f) + (1-\alpha)u_k \\ &=& (\alpha T + (1-\alpha)I)u_k + \alpha f \equiv T_\alpha u_k + \alpha f. \end{array}\right. \tag{4.9.4}$$

is called an Extrapolation method of (4.9.2).

Remark 4.9.1 *$T_\alpha = \alpha T + (1-\alpha)I$ is a new iterative matrix ($T_1 = T$). T_α arises from the decomposition $A = \frac{1}{\alpha}M - (N + (\frac{1}{\alpha} - 1)M)$.*

Theorem 4.9.2 *If (4.9.2) is symmetrizable and T has the eigenvalues satisfying $\mu_1 \le \mu_2 \le \cdots \le \mu_n < 1$, then it holds for $\alpha^* = \frac{2}{2-\mu_1-\mu_2} > 0$ that*

$$1 > \rho(T_{\alpha^*}) = \frac{\mu_n - \mu_1}{2 - \mu_1 - \mu_n} = \min_\alpha \rho(T_\alpha). \tag{4.9.5}$$

Proof: Eigenvalues of T_α are $\alpha\mu_i + (1-\alpha) = 1 + \alpha(\mu_i - 1)$. Consider the problem

$$\begin{array}{rl} & \min\limits_\alpha \max\limits_i |1 + \alpha(u_i - 1)| = \text{min!} \\ \Longleftrightarrow & |1 + \alpha(\mu_n - 1)| = |1 + \alpha(\mu_1 - 1)|, \\ \Longleftrightarrow & 1 + \alpha(\mu_n - 1) = \alpha(1 - \mu_n) - 1 \text{ (otherwise } \mu_1 = \mu_n). \end{array}$$

This implies $\alpha = \alpha^* = \frac{2}{2-\mu_1-\mu_n}$, then $1 + \alpha^*(\mu_n - 1) = \frac{\mu_n - \mu_1}{2-\mu_1-\mu_n}$. ■

From (4.9.2) and (4.9.4) follows that

$$u_k = \sum_{i=0}^{k} a_{ki} x_i, \text{ and } \sum_{i=0}^{k} a_{ki} = 1$$

with suitable a_{ki}. Hence, we have the following idea:

Find a sequence $\{a_{ki}\}$, $k = 1, 2, \ldots$, $i = 0, 1, 2, \ldots, k$ and $\sum\limits_{i=0}^{k} a_{ki} = 1$ such that

$$u_k = \sum_{i=0}^{k} a_{ki} x_i, \; u_0 = x_0 \tag{4.9.6}$$

is a good approximation of x^* ($Ax^* = b$). Hereby the cost of computation of u_k should not be more expensive than x_k.

Error: Let

$$e_k = x_k - x^*, e_k = T^k e_0, \; e_0 = x_0 - x^* = u_0 - x^* = d_0. \tag{4.9.7}$$

Hence,

$$\begin{array}{rcll} d_k &=& u_k - x^* = \sum\limits_{i=0}^{k} a_{ki}(x_i - x^*) & (4.9.8) \\ &=& \sum\limits_{i=0}^{k} a_{ki} T^i e_0 = (\sum\limits_{ki}^{k} a_{ki} T^i) e_0 & \\ &=& \mathcal{P}_k(T) e_0 = \mathcal{P}_k(T) d_0, & \end{array}$$

where

$$\mathcal{P}_k(\lambda) = \sum_{i=0}^{k} a_{ki}\lambda^i \tag{4.9.9}$$

is a polynomial in λ with $\mathcal{P}_k(1) = 1$.
Problem: Find $\mathcal{P}_k$ such that $\rho(\mathcal{P}_k(T))$ is possible small.

Remark 4.9.2 *Let $\|x\|_W = \|Wx\|_2$. Then*

$$\begin{aligned} \|T\|_W &= \max_{x\neq 0} \frac{\|Tx\|_W}{\|x\|_W} \\ &= \max_{x\neq 0} \frac{\|WTW^{-1}Wx\|_2}{\|Wx\|_2} \\ &= \|WTW^{-1}\|_2 = \rho(T), \end{aligned}$$

because WTW^{-1} is symmetric. We take $\|\cdot\|_W$-norm on both sides of (4.9.8) and have

$$\begin{aligned} \|d_k\|_W &\leq \|\mathcal{P}_k(T)\|_W\|d_0\|_W = \|W\mathcal{P}_k(T)W^{-1}\|_2\|d_0\|_2 \\ &= \|\mathcal{P}_k(WTW^{-1})\|_2\|d_0\|_W = \rho(\mathcal{P}_k(T))\|d_0\|_W. \end{aligned} \tag{4.9.10}$$

Replacement problem: Let $1 > \mu_n \geq \cdots \geq \mu_1$ be the eigenvalues of T. Determine

$$\min\left[\{\max|\mathcal{P}_k(\lambda)| : \mu_1 \leq \lambda \leq \mu_n\} : \deg(\mathcal{P}_k) \leq k,\ \mathcal{P}_k(1) = 1\right]. \tag{4.9.11}$$

Solution of (4.9.11): The replacement problem (4.6.35)

$$\max\{|\mathcal{P}_k(\lambda)| : 0 < a \leq \lambda \leq b\} = \min!,\ \mathcal{P}_k(0) = 1$$

has the solution

$$Q_k(t) = T_k(\frac{2t-b-a}{b-a}) \Big/ T_k(\frac{b+a}{a-b}).$$

Substituting $t \to 1-\lambda$, $\lambda \to 1-t$, $(\mu_1, \mu_n) \to (1-\mu_n, 1-\mu_1)$, the problem (4.9.11) can be transformed to the problem (4.6.35). Hence, the solution of (4.9.11) is given by

$$Q_k(t) = T_k(\frac{2t-\mu_1-\mu_n}{\mu_1-\mu_n}) \Big/ T_k(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}). \tag{4.9.12}$$

Write $Q_k(t) := \sum_{i=0}^{k} a_{ki}t^i$. Then we have

$$u_k = \sum_{i=0}^{k} a_{ki}x_i,$$

which is called the **optimal Chebychev semi-iterative method**.
Effective Computation of u_k: Using recursion of T_k as in (4.6.36), we get

$$T_0(t) = 1,\ T_1(t) = t,\ T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t).$$

Transforming $T_k(t)$ to the form of $Q_k(t)$ as in (4.9.12) we get

$$Q_0(t) = 1,\ Q_1(t) = \frac{2t - \mu_1 - \mu_n}{2 - \mu_1 - \mu_n} = pt + (1 - p) \tag{4.9.13a}$$

and

$$Q_{k+1}(t) = [pt + (1 - p)]c_{k+1}Q_k(t) + (1 - c_{k+1})\, Q_{k-1}(t), \tag{4.9.13b}$$

where

$$p = \frac{2}{2 - \mu_1 - \mu_n},\ c_{k+1} = \frac{2T_k(1/r)}{rT_{k+1}(1/r)} \text{ and } r = \frac{\mu_1 - \mu_n}{2 - \mu_1 - \mu_n}. \tag{4.9.14}$$

Claim: (4.9.13b)

$$\begin{aligned}
Q_{k+1}(t) &= T_{k+1}\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big) \Big/ T_{k+1}\Big(\frac{1}{r}\Big) \\
&= \frac{1}{T_{k+1}(1/r)}\Big[2\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big)T_k\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big) - T_{k-1}\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big)\Big] \\
&= \frac{2T_k(1/r)}{rT_{k+1}(1/r)} r\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big)\frac{T_k\big(\frac{2t-\mu_1-\mu_n}{\mu_1-\mu_n}\big)}{T_k(1/r)} - \frac{T_{k-1}\big(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}\big)}{T_{K+1}\big(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}\big)}\frac{T_{k-1}\big(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}\big)}{T_{k-1}(1/r)} \\
&= C_{k+1}[pt + (1 - p)]Q_k(t) - [1 - C_{k+1}]Q_{k-1}(t),
\end{aligned}$$

since

$$r\Big(\frac{2t - \mu_1 - \mu_n}{\mu_1 - \mu_n}\Big) = \frac{2t - \mu_1 - \mu n}{2 - \mu_1 - \mu_n} = pt + (1 - p)$$

and

$$\begin{aligned}
1 - C_{k+1} &= 1 - \frac{2T_k(1/r)}{rT_{k+1}(1/r)} = \frac{rT_{k+1}(1/r) - 2T_k(1/r)}{rT_{k+1}(1/r)} \\
&= \frac{-rT_{k-1}(1/r)}{rT_{k+1}(1/r)} = \frac{-T_{k-1}(1/r)}{T_{k+1}(1/r)}.
\end{aligned}$$

Recursion for u_k:

$$\begin{aligned}
d_{k+1} &= Q_{k+1}(T)d_0 = (pT + (1 - p)I)c_{k+1}Q_k(T)d_0 + (1 - c_{k+1})Q_{k-1}(T)d_0, \\
x^* &= (pT + (1 - p)I)c_{k+1}x^* + (1 - c_{k+1})x^* + p(I - T)x^*c_{k+1}.
\end{aligned}$$

Adding above two equations together we get

$$\begin{aligned}
u_{k+1} &= [pT + (1 - p)I]c_{k+1}u_k + (1 - c_{k+1})u_{k-1} + c_{k+1}pf \\
&= c_{k+1}p\,\{Tu_k + f - u_k\} + c_{k+1}u_k + (1 - c_{k+1})\, u_{k-1}.
\end{aligned}$$

Then we obtain the optimal Chebychev semi-iterative Algorithm.

Algorithm 4.9.1 (Optimal Chebychev semi-iterative Algorithm)

$$\begin{aligned}
&\textit{Let } r = \tfrac{\mu_1-\mu_n}{2-\mu_1-\mu_n},\ p = \tfrac{2}{2-\mu_1-\mu_n},\ c_1 = 2 \\
&u_0 = x_0, \\
&u_1 = p(Tu_0 + f) + (1 - p)u_0 \\
&\textit{For } k = 1, 2, \cdots, \\
&\quad u_{k+1} = c_{k+1}\,[p(Tu_k + f) + (1 - p)u_k] + (1 - c_{k+1})\, u_{k-1}, \\
&\quad c_{k+1} = (1 - r^2/4\, c_k)^{-1}.
\end{aligned} \tag{4.9.15}$$

Remark 4.9.3 *Here u_{k+1} can be rewritten as the three terms recursive formula with two parameters as in (4.7.5):*

$$\begin{aligned} u_{k+1} &= c_{k+1}\left[p\left(Tu_k+f\right)+(1-p)u_k\right]+\left(1-c_{k+1}\right)u_{k-1} \\ &= c_{k+1}\left[pM^{-1}\left((M-A)u_k+b\right)+(1-p)u_k\right]+u_{k-1}-c_{k+1}u_{k-1} \\ &= c_{k+1}\left[u_k+pM^{-1}\left(b-Au_k\right)-u_{k-1}\right]+u_{k-1} \\ &= c_{k+1}\left[u_k+pz_k-u_{k-1}\right]+u_{k-1}, \end{aligned}$$

where $Mz_k = b - Au_k$. ■

Recursion for c_k: Since

$$c_1 = \frac{2t_0}{rT_1(1/r)} = \frac{2}{r\cdot\frac{1}{r}} = 2,$$

thus

$$T_{k+1}\left(\frac{1}{r}\right) = \frac{2}{r}T_k\left(\frac{1}{r}\right) - T_{k-1}\left(\frac{1}{r}\right) \quad \text{(from (4.6.36))}.$$

It follows

$$\frac{1}{c_{k+1}} = \frac{rT_{k+1}\left(\frac{1}{r}\right)}{2T_k\left(\frac{1}{r}\right)} = 1-\frac{r^2}{4}\left[\frac{2T_{k-1}\left(\frac{1}{r}\right)}{rT_k\left(\frac{1}{r}\right)}\right] = 1-\frac{r^2}{4}c_k.$$

Then we have

$$c_{k+1} = \frac{1}{\left(1-\left(r^2/4\right)c_k\right)} \quad \text{with} \quad r = \frac{\mu_1-\mu_n}{2-\mu_1-\mu_n}. \tag{4.9.16}$$

Error estimate: It holds

$$\|u_k - x^*\|_W \le \left|T_k\left(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}\right)\right|^{-1}\|u_0-x^*\|_W. \tag{4.9.17}$$

Proof: From (4.9.10) and (4.9.12) we have

$$\begin{aligned} \|d_k\|_W &= \|Q_k(T)d_0\|_W \le \rho\left(Q_k(T)\right)\|d_0\|_W \\ &\le \max\left\{|Q_k(\lambda)| : \mu_1\le\lambda\le\mu_n\right\}\|d_0\|_W \\ &\le \left|T_k\left(\frac{2-\mu_1-\mu_n}{\mu_1-\mu_n}\right)\right|^{-1}\|d_0\|_W. \end{aligned}$$

■

We want to estimate the quantity $q_k := |T_k(1/r)|^{-1}$ (see also Lemma 4.6.9). From (4.6.37) we have

$$\begin{aligned} T_k\left(\frac{1}{r}\right) &= \frac{1}{2}\left[\left(\frac{1+\sqrt{1-r^2}}{r}\right)^k + \left(\frac{1-\sqrt{1-r^2}}{r}\right)^k\right] \\ &= \frac{1}{2}\left[\frac{(1+\sqrt{1-r^2})^k+(1-\sqrt{1-r^2})^k}{(r^2)^{k/2}}\right] \\ &= \frac{1}{2}\left[\frac{(1+\sqrt{1-r^2})^k+(1-\sqrt{1-r^2})^k}{\left[(1+\sqrt{1-r^2})(1-\sqrt{1-r^2})\right]^{k/2}}\right] \\ &= \frac{1}{2}\left(c^{k/2}+c^{-k/2}\right) \ge \frac{1}{2c^{k/2}}\ , \end{aligned}$$

μ_n	k	q_4	j	j'	q_8	j	j'
0.8	5	0.0426	8	14	9.06(-4)	17–18	31
0.9	10	0.1449	9–10	18	1.06(-2)	22–23	43
0.95	20	0.3159	11–12	22	5.25(-2)	29–30	57
0.99	100	0.7464	14–15	29	3.86(-1)	47	95

Table 4.3: Convergence rate of q_k where $j : \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^j \approx q_4,\ q_8$ and $j' : \mu_n^{j'} \approx q_4,\ q_8$.

where $c = \frac{1-\sqrt{1-r^2}}{1+\sqrt{1-r^2}} < 1$. Thus $q_k \leq 2c^{k/2}$. Rewrite the eigenvalues of $I-T$ as $\lambda_i = 1-\mu_i$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. Then

$$r = \frac{\mu_n - \mu_1}{2-\mu_1-\mu_n} = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{\kappa - 1}{\kappa + 1}\ , \quad \kappa = \frac{\lambda_1}{\lambda_n}$$

Thus, from $c = \frac{1-\sqrt{1-r^2}}{1+\sqrt{1-r^2}} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ follows

$$q_k \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k. \tag{4.9.18}$$

That is, after k steps of the Chebychev semi-iterative method the residual $\|u_k - x^*\|_W$ is reduced by a factor $2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$ from the original residual $\|u_0 - x^*\|_W$.

If $\mu_{\min} = \mu_1 = 0$, then $q_k = T_k\left(\frac{2-\mu_n}{\mu_n}\right)^{-1}$. Table 4.3 shows the convergence rate of the quantity q_k.

All above statements are true, if we replace μ_n by μ_n' ($\mu_n' \geq \mu_n$) and μ_1 by μ_1' ($\mu_1' \leq \mu_1$), because λ is still in $[\mu_1', \mu_n']$ for all eigenvalue λ of T.

Example 4.9.2 *Let $1 > \rho = \rho(T)$. If we set $\mu_n' = \rho, \mu_1' = -\rho$, then p and r defined in (4.9.14) become $p = 1$ and $r = \rho$, respectively. Algorithm 4.9.1 can be simplified by*

$$\begin{aligned} &u_0 = x_0, \\ &u_1 = Tu_0 + f, \\ &u_{k+1} = c_{k+1}(Tu_k + f) + (1 - c_{k+1})u_{k-1}, \\ &c_{k+1} = \left(1 - \left(\rho^2/4\right)c_k\right)^{-1} \quad \textit{with } c_1 = 2. \end{aligned} \tag{4.9.19}$$

■

Also, Algorithm 4.9.1 can be written by the form of (4.9.19), by replacing T by $T_{\alpha^*} = T_p = (pT + (1-p)I)$ and it leads to

$$u_{k+1} = c_{k+1}\left(T_p u_k + f\right) + \left(1 - c_{k+1}\right)u_{k-1}. \tag{4.9.20}$$

Here $p\mu_1 + (1-p) = \frac{\mu_1 - \mu_n}{2-\mu_1-\mu_n}$ and $p\mu_n + (1-p) = \frac{\mu_n - \mu_1}{2-\mu_1-\mu_n}$ are eigenvalues of T_p.

Remark 4.9.4 *(i) In (4.9.15) it holds ($r = \rho$)*

$$c_2 > c_3 > c_4 > \cdots, \ \textit{and} \ \lim_{k\to\infty} c_k = \frac{2}{1+\sqrt{1-r^2}}. \quad \textit{(Exercise!)}$$

(ii) If T is symmetric, then by (4.9.12) we get

$$\begin{aligned} \|Q_k(T)\|_2 &= \max\{|Q_k(\mu_i)| : \mu_i \text{ is an eigenvalue of } T\} \\ &\le \max\{|Q_k(\lambda)| : -\rho \le \lambda \le \rho\} \\ &= \left|T_k\left(\frac{1}{\rho}\right)\right|^{-1}, \quad (\rho = \rho(T)). \\ &= \frac{1}{c^{k/2} + c^{-k/2}} = \frac{(\omega_b - 1)^{k/2}}{1 + (\omega_b - 1)^k}, \end{aligned} \tag{4.9.21}$$

where $c = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \omega_b - 1$ *with* $\omega_b = \frac{2}{1+\sqrt{1-\rho^2}}$.

4.9.1 Connection with SOR Method

Recall

(i) The SOR method solves linear system $Ax = b$ (standard decomposition $A = I - L - R$):

$$\begin{aligned} x^{(i+1)} &= (I - \omega L)^{-1}((1-\omega)I + \omega R)x^{(i)} + \omega(I - \omega L)^{-1}b \\ &= L_\omega x^{(i)} + \omega(I - \omega L)^{-1}b, \qquad 0 < \omega < 2 \end{aligned} \tag{4.9.22}$$

(ii) $A = I - L - R$ is called 2–consistly ordered, if the eigenvalues of $\alpha L + \alpha^{-1}R$ are independent of α,

(iii) (Theorem) $A = I - L - R$ and A is 2–consistly ordered. If A has real eigenvalues and $\rho(L + R) < 1$, then it holds

$$\omega_b - 1 = \rho\left(L_{\omega_b}\right) < \rho(L_\omega), \quad \omega \ne \omega_b, \tag{4.9.23}$$

where $\omega_b = \frac{2}{1+\sqrt{1-\rho^2(L+R)}}$.

Consider (4.9.1) again

$$x = Tx + f, \quad A = M - N, \quad T = M^{-1}N, \quad f = M^{-1}b.$$

Assume that

$$\text{all eigenvalues of } T \text{ are real and } \rho(T) < 1. \tag{4.9.24}$$

Then the following linear system (of order $2n$) is equivalent to (4.9.1).

$$\begin{cases} x = Ty + f, \\ y = Tx + f. \end{cases} \tag{4.9.25}$$

That is, if x^* solves (4.9.1), then $\begin{bmatrix} x^* \\ x^* \end{bmatrix}$ solves (4.9.25), reversely, if $z^* = \begin{bmatrix} {z_1}^* \\ {z_2}^* \end{bmatrix}$ solves (4.9.25), then ${z_1}^* = {z_2}^*$ solves (4.9.1). Because ${z_1}^* - {z_2}^* = -T({z_1}^* - {z_2}^*)$ and -1 is not an eigenvalue of T, so ${z_1}^* = {z_2}^*$. Let

$$z = \begin{bmatrix} x \\ y \end{bmatrix}, \quad J = \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix}, \quad h = \begin{bmatrix} f \\ f \end{bmatrix}.$$

Then (4.9.25) can be written as

$$z = Jz + h \tag{4.9.26}$$

and $I - J$ is 2–consistly ordered. Applying SOR method to (4.9.26) we get

$$J = L + R := \begin{bmatrix} 0 & 0 \\ T & 0 \end{bmatrix} + \begin{bmatrix} 0 & T \\ 0 & 0 \end{bmatrix}$$

and

$$(I - \omega L)z_{i+1} = ((1-\omega)I + \omega R)z_i + \omega h. \tag{4.9.27}$$

Let $z_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$. Then we have

$$\begin{bmatrix} I & 0 \\ -\omega T & I \end{bmatrix} \begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} (1-\omega)I & \omega T \\ 0 & (1-\omega)I \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} \omega f \\ \omega f \end{bmatrix},$$

hence

$$x_{i+1} = (1-\omega)x_i + \omega T y_i + \omega f = \omega\{Ty_i + f - x_i\} + x_i, \tag{4.9.28a}$$
$$y_{i+1} = \omega T x_{i+1} + (1-\omega)y_i + \omega f = \omega\{Tx_{i+1} + f - y_i\} + y_i, \tag{4.9.28b}$$

The optimal value ω_b for (4.9.27) is given by

$$\omega_b = \frac{1}{1 + \sqrt{1 - \rho^2(J)}} .$$

Lemma 4.9.3 *It holds $\sigma(J) = \sigma(T) \cup \{-\sigma(T)\}$, where $\sigma(T)$ = spectrum of T. Especially $\rho(T) = \rho(J)$.*

Proof: Let $\lambda \in \sigma(T)$. There exists $x \neq 0$ with $Tx = \lambda x$. Then

$$J \begin{bmatrix} x \\ x \end{bmatrix} = \lambda \begin{bmatrix} x \\ x \end{bmatrix} \quad \text{and} \quad J \begin{bmatrix} x \\ -x \end{bmatrix} = -\lambda \begin{bmatrix} x \\ -x \end{bmatrix}.$$

Thus we have $\sigma(J) \supset \sigma(T) \cup \{-\sigma(T)\}$. On the other hand, from $J^2 = \begin{bmatrix} T^2 & 0 \\ 0 & T^2 \end{bmatrix}$ follows that if λ is an eigenvalue of J, then $\lambda^2 = \mu^2$ for one $\mu \in \sigma(T)$, so $\lambda = \mu$ or $-\mu$. Thus

$$\sigma(J) \subset \sigma(T) \cup \{-\sigma(T)\}.$$

■

We then have

$$\omega_b = \frac{2}{1 + \sqrt{1 - \rho^2(T)}}, \quad \rho(L\omega_b) = \omega_b - 1 = \frac{1 - \sqrt{1 - \rho^2(T)}}{1 + \sqrt{1 - \rho^2(T)}} . \tag{4.9.29}$$

4.9.2 Practical Performance

$$\begin{array}{cccccccc} x_0, & y_0, & x_1, & y_1, & x_2, & y_2, & \cdots \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \\ \zeta_0, & \zeta_1, & \zeta_2, & \zeta_3, & \zeta_4, & \zeta_5, & \cdots \end{array}$$

$$\zeta_{2i} = x_i, \quad \zeta_{2i+1} = y_i, \quad i = 0, 1, 2, \ldots$$

Then (4.9.28) can be written as

$$\zeta_{i+1} = \omega_b\{T\zeta_i + f - \zeta_{i-1}\} + \zeta_{i-1}, \quad i = 1, 2, \cdots \tag{4.9.30}$$

with $\zeta_0 = x_0$ and $\zeta_1 = y_0 = Tx_0 + f$. Comparing (4.9.30) with (4.9.19) we get

$$u_{i+1} = c_{i+1}\{Tu_i + f - u_{i-1}\} + u_{i-1}, \quad i = 1, 2, \cdots . \tag{4.9.31}$$

Since c_i converges to ω_b, the optimal Chebychev acceleration method is referred to as a variant SOR method.

Error estimate of (4.9.30): Write (4.9.30) as

$$\begin{aligned} \zeta_{k+1} &= \omega_b\{T\zeta_k + f - \zeta_{k-1}\} + \zeta_{k-1}, \\ \zeta_0 &= x_0, \\ \zeta_1 &= T\zeta_0 + f. \end{aligned}$$

Let

$$\varepsilon_k = \zeta_k - x^*. \quad (Ax^* = b) \tag{4.9.32}$$

Then we have

$$\begin{aligned} \varepsilon_0 &= \zeta_0 - x^*, \\ \varepsilon_1 &= T\varepsilon_0, \\ \varepsilon_{k+1} &= \omega_b T\varepsilon_k + (1 - \omega_b)\varepsilon_{k-1}. \end{aligned}$$

Since $x^* = \omega_b\{Tx^* + f - x^*\} + x^*$, it follow that

$$\varepsilon_k = r_k(T)\varepsilon_0, \tag{4.9.33}$$

where $r_k(x)$ is a polynomial of degree $\leq k$, and

$$\begin{aligned} r_0 &= 1, \\ r_1(t) &= t, \\ r_{k+1}(t) &= \omega_b t r_k(t) + (1 - \omega_b) r_{k-1}(t). \end{aligned} \tag{4.9.34}$$

Either solve this difference equation or reduce to Chebychev polynomials of 2nd kind.

$$\begin{aligned} s_{k+1}(t) &= 2ts_k(t) - s_{k-1}(t), \\ s_0(t) &= 1, \\ s_1(t) &= 2t. \end{aligned}$$

In fact $s_k(\cos\theta) = \sin((k+1)\theta)/\sin\theta$. One can estimate $\|r_k(T)\|$ (see Varga p.146) by: Let T be Hermitian. Then

$$\begin{aligned}\|r_k(T)\| &= \max\{|r_k(\mu_i)| : \mu_i \text{ is an eigenvalue of } T\}\\ &= \max\{|r_k(\mu)| : -\rho(T) \le \mu \le \rho(T)\}\\ &= (\omega_b - 1)^{k/2}\left\{1 + k\sqrt{1-\rho^2(T)}\right\}\end{aligned}$$

This implies

$$\lim_{k\to\infty} \|r_k(T)\|^{1/k} = \sqrt{\omega_b - 1}\ .$$

From (4.9.21) follows that

$$\lim_{k\to\infty} \|Q_k(T)\|^{1/k} = \sqrt{\omega_b - 1}\ .$$

4.10 GCG-type Methods for Nonsymmetric Linear Systems

Recall: A is s.p.d. Consider the quadratic functional

$$F(x) = \frac{1}{2}x^T Ax - x^T b \tag{4.10.1}$$

$$Ax^* = b \Longleftrightarrow \min_{x\in\mathbb{R}^n} F(x) = F(x^*)$$

Consider

$$\varphi(x) = \frac{1}{2}(b-Ax)^T A^{-1}(b-Ax) = F(x) + \frac{1}{2}b^T A^{-1} b, \tag{4.10.2}$$

where $\frac{1}{2}b^T A^{-1}b$ is a constant. Then

$$Ax^* = b \iff \varphi(x^*) = \min_{x\in\mathbb{R}^n}\varphi(x) = [\min_{x\in\mathbb{R}^n} F(x)] + \frac{1}{2}b^T A^{-1}b$$

CG-method:

Given $x_0, r_0 = p_0 = b - Ax_0$
for $k = 0, 1, \ldots$
$\alpha_k = r_k^T p_k / p_k^T A p_k$,
$x_{k+1} = x_k + \alpha_k p_k$,
$r_{k+1} = r_k - \alpha_k A p_k$ $(\equiv b - Ax_{k+1})$
$p_{k+1} = r_{k+1} + \beta_k p_k$
$\beta_k = -r_{k+1}^T A p_k / p_k^T A p_k (= r_{k+1}^T r_{k+1}/r_k^T r_k)$
end for

Numerator: $r_{k+1}^T((r_k - r_{k+1})/\alpha_k) = (-r_{k+1}^T r_{k+1})/\alpha_k$
Denominator: $p_k^T A p_k = (r_k^T + \beta_{k-1}p_{k-1}^T)((r_k - r_{k+1})/\alpha_k) = (r_k^T r_k)/\alpha_k$.

Remark 4.10.1 *CG method does not need to compute any parameters. It only needs matrix vector and inner product of vectors. Hence it can not destroy the sparse structure of the matrix A.*

The vectors r_k and p_k generated by CG-method satisfy:

$$\begin{aligned} p_i^T r_k &= (p_i, r_k) = 0, \quad i < k \\ r_i^T r_j &= (r_i, r_j) = 0, \quad i \neq j \\ p_i^T A p_j &= (p_i, Ap_j) = 0, \quad i \neq j \end{aligned}$$

$x_{k+1} = x_0 + \sum_{i=0}^{k} \alpha_i p_i$ minimizes $F(x)$ over $x = x_0 + < p_0, \cdots, p_k >$.

4.10.1 GCG method(Generalized Conjugate Gradient)

GCG method is developed to minimize the residual of the linear equation under some special functional. In conjugate gradient method we take

$$\varphi(x) = \frac{1}{2}(b - Ax)^T A^{-1}(b - Ax) = \frac{1}{2} r^T A^{-1} r = \frac{1}{2}\|r\|^2_{A^{-1}},$$

where $\|x\|_{A^{-1}} = \sqrt{x^T A^{-1} x}$.

Let A be a unsymmetric matrix. Consider the functional

$$f(x) = \frac{1}{2}(b - Ax)^T P(b - Ax),$$

where P is s.p.d. Thus $f(x) > 0$, unless $x^* = A^{-1}b \Rightarrow f(x^*) = 0$, so x^* minimizes the functional $f(x)$.

Different choices of P:

(i) $P = A^{-1}$ (A is s.p.d.) $\Rightarrow$ CG method (classical)

(ii) $P = I \Rightarrow$ GCR method (Generalized Conjugate residual).

$$f(x) = \frac{1}{2}(b - Ax)^T(b - Ax) = \frac{1}{2}\|r\|_2^2$$

Here $\{r_i\}$ forms A-conjugate.

(iii) Consider $M^{-1}Ax = M^{-1}b$. Take $P = M^T M > 0 \Rightarrow$ GCGLS method (Generalized Conjugate Gradient Least Square).

(iv) Similar to (iii), take $P = (A + A^T)/2$ (note: P is not positive definite) and $M = (A + A^T)/2$ we get GCG method (by Concus, Golub and Widlund). In general, P is not necessary to be taken positive definite, but it must be symmetric ($P^T = P$). Therefore, the minimality property does not hold.

Let

$$(x, y)_o = x^T P y \Longrightarrow (x, y)_o = (y, x)_o.$$

Algorithm 4.10.1 (GCG method)

Given $x_0,\ r_0 = p_0 = b - Ax_0$

for $k = 0, 1, \cdots$

$$\alpha_k = (r_k, Ap_k)_o/(Ap_k, Ap_k)_o \tag{4.10.3a}$$

$$x_{k+1} = x_k + \alpha_k p_k \tag{4.10.3b}$$

$$r_{k+1} = r_k - \alpha_k Ap_k\ (\equiv b - Ax_{k+1}) \tag{4.10.3c}$$

$$\beta_i^{(k)} = -(Ar_{k+1}, Ap_i)_o/(Ap_i, Ap_i)_o, \quad i = 0, 1, \cdots, k \tag{4.10.3d}$$

$$p_{k+1} = r_{k+1} + \sum_{i=0}^{k} \beta_i^{(k)} p_i \tag{4.10.3e}$$

end for

In GCG method, the choice of $\{\beta_i^{(k)}\}_{i=1}^k$ satisfy:

$$(r_{k+1}, Ap_i)_o = 0, \quad i \le k \tag{4.10.4a}$$

$$(r_{k+1}, Ar_i)_o = 0, \quad i \le k \tag{4.10.4b}$$

$$(Ap_i, Ap_j)_o = 0, \quad i \ne j \tag{4.10.4c}$$

Theorem 4.10.1 *$x_{k+1} = x_0 + \sum_{i=0}^k \alpha_k p_i$ minimizes $f(x) = \frac{1}{2}(b - Ax)^T P(b - Ax)$ over $x = x_0 + < p_0, \cdots, p_k >$, where P is s.p.d.*

(The proof is the same as that of classical CG method).

If P is indefinite, which is allowed in GCG method, then the minimality property does not hold. x_{k+1} is the critical point of $f(x)$ over $x = x_0 + < p_0, \cdots, p_k >$.

Question: Can the GCG method break down? i.e., Can α_k in GCG method be zero? Consider the numerator of α_k:

$$\begin{aligned} (r_k, Ap_k) &= (r_k, Ar_k)_o \qquad \text{[by (4.10.3e) and (4.10.4a)]} \\ &= r_k^T PAr_k \\ &= r_k^T A^T Pr_k \qquad \text{[Take transpose]} \\ &= r_k^T \tfrac{(PA + A^T P)}{2} r_k. \end{aligned} \tag{4.10.5}$$

From (4.10.5), if $(PA + A^T P)$ is positive definite, then $\alpha_k \ne 0$ unless $r_k = 0$. Hence if the matrix A satisfies $(PA + A^T P)$ positive definite, then GCG method can not break down.

From GCG method, r_k and p_k can be rewritten by

$$r_k = \psi_k(A) r_0, \tag{4.10.6a}$$

$$p_k = \varphi_k(A) r_0, \tag{4.10.6b}$$

where ψ_k and φ_k are polynomials of degree $\le k$ with $\psi_k(0) = 1$ [by (4.10.3c), (4.10.3e)]. From (4.10.6a), (4.10.6b) and (4.10.4b) follows that

$$(r_{k+1}, A^{i+1} r_0)_o = 0, \qquad i = 0, 1, \cdots, k. \tag{4.10.7}$$

From (4.10.6a), (4.10.6b) and (4.10.3d), the numerator of $\beta_i^{(k)}$ can be expressed by

$$(Ar_{k+1}, Ap_i)_o = r_{k+1}^T A^T PAp_i = r_{k+1}^T A^T PA\varphi_i(A) r_0. \tag{4.10.8}$$

If A^TP can be expressed by

$$A^TP = P\theta_s(A), \tag{4.10.9}$$

where θ_s is some polynomial of degree s. Then (4.10.8) can be written by

$$\begin{aligned}(Ar_{k+1}, Ap_i)_o &= r_{k+1}^T A^T P A\varphi_i(A)r_0 \\ &= r_{k+1}^T P\theta_s(A)A\varphi_i(A)r_0 \\ &= (r_{k+1}, A\theta_s(A)\varphi_i(A)r_0)_o.\end{aligned} \tag{4.10.10}$$

From (4.10.7) we know that if $s+i \le k$, then (4.10.10) is zero, i.e.,$(Ar_{k+1}, Ap_i)_o = 0$. Hence $\beta_i^{(k)} = 0,\ i = 0, 1, \cdots, k-s$. But only in the special case s will be small. For instance,

(i) In classical CG method, A is s.p.d, P is taking by A^{-1}. Then $A^TP = AA^{-1} = I = A^{-1}A = A^{-1}\theta_1(A)$, where $\theta_1(x) = x, s = 1$. So, $\beta_i^{(k)} = 0$, for all $i+1 \le k$, it is only $\beta_k^{(k)} \neq 0$.

(ii) Concus, Golub and Widlund proposed GCG method, it solves $M^{-1}Ax = M^{-1}b$. (A: unsymmetric), where $M = (A+A^T)/2$ and $P = (A+A^T)/2$ (P may be indefinite).

- Check condition (4.10.9):

$$(M^{-1}A)^TP = A^TM^{-1}M = A^T = M(2I - M^{-1}A) = P(2I - M^{-1}A).$$

Then

$$\theta_s(M^{-1}A) = 2I - M^{-1}A,$$

where $\theta_1(x) = 2 - x,\ s = 1$. Thus $\beta_i^{(k)} = 0,\ i = 0, 1, \cdots, k-1$. Therefore we only use r_{k+1} and p_k to construct p_{k+1}.

- Check condition $A^TP + PA$:

$$(M^{-1}A)^TM + MM^{-1}A = A^T + A \quad \text{indefinite}$$

The method can possibly break down.

(iii) The other case $s = 1$ is BCG (BiCG) (See next paragraph).

Remark 4.10.2 *Except the above three cases, the degree s is usually very large. That is, we need to save all directions p_i $(i = 0, 1, \cdots, k)$ in order to construct p_{k+1} satisfying the conjugate orthogonalization condition (4.10.4c). In GCG method, each iteration step needs to save $2k+5$ vectors (x_{k+1}, r_{k+1}, p_{k+1}, $\{Ap_i\}_{i=0}^k$, $\{p_i\}_{i=0}^k$), $k+3$ inner products (Here k is the iteration number). Hence, if k is large, then the space of storage and the computation cost can become very large and can not be acceptable. So, GCG method, in general, has some practical difficulty. Such as GCR, GMRES (by SAAD) methods, they preserve the optimality ($p > 0$), but it is too expensive (s is very large).*

Modification:

(i) Restarted: If GCG method does not converge after $m+1$ iterations, then we take x_{k+1} as x_0 and restart GCG method. There are at most $2m+5$ saving vectors.

(ii) Truncated: The most expensive step of GCG method is to compute $\beta_i^{(k)},\ i = 0, 1, \cdots, k$ so that p_{k+1} satisfies (4.10.4c). We now release the condition (4.10.4c) to require that p_{k+1} and the nearest m direction $\{p|i\}_{i=k-m+1}^k$ satisfy the conjugate orthogonalization condition.

4.10.2 BCG method (A: unsymmetric)

BCG method is similar to the CG method, it does not need to save the search direction. But the norm of the residual produced by BCG method does not preserve the minimal property.

Solve $Ax = b$ by considering $A^T y = c$ (phantom). Let

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & A^T \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} b \\ c \end{pmatrix}.$$

Consider

$$\tilde{A}\tilde{x} = \tilde{b}.$$

Take $P = \tilde{A}^{-T} Z$ ($P = P^T$) with $Z = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$. This implies

$$\tilde{A}^T Z = Z\tilde{A} \quad \text{and} \quad \tilde{A}^T P = P\tilde{A}.$$

From (4.10.9) we know that $s = 1$ for $\tilde{A}\tilde{x} = \tilde{b}$. Hence it only needs to save one direction p_k as in the classical CG method.

Algorithm 4.10.2 (Apply GCG method to $\tilde{A}\tilde{x} = \tilde{b}$)

Given $x_0 = \begin{pmatrix} x_0 \\ \hat{x_0} \end{pmatrix}$, $\tilde{p}_0 = \tilde{r}_0 = \tilde{b} - \tilde{A}\tilde{x}_0 = \begin{pmatrix} r_0 \\ \hat{r_0} \end{pmatrix}$.
for $k = 0, 1, \ldots$
$\alpha_k = (\tilde{r}_k, \tilde{A}\tilde{p}_k)_o / (\tilde{A}\tilde{p}_k, \tilde{A}\tilde{p}_k)_o$,
$\tilde{x}_{k+1} = \tilde{x}_k + \alpha_k \tilde{p}_k$,
$\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k \tilde{A}\tilde{p}_k, \quad \tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_k \tilde{p}_k$
$\beta_k = -(\tilde{A}\tilde{r}_{k+1}, \tilde{A}\tilde{p}_k)_o / (\tilde{A}\tilde{p}_k, \tilde{A}\tilde{p}_k)_o$.
end for

Algorithm 4.10.3 (Simplification (BCG method))

Given $x_0, p_0 = r_0 = b - Ax_0$
Choose $\hat{r}_0, \hat{p}_0 = \hat{r}_0$
for $k = 0, 1, \ldots$
$\alpha_k = (\hat{r}_k, r_k)/(\hat{p}_k, Ap_k)$,
$x_{k+1} = x_k + \alpha_k p_k$,
$r_{k+1} = r_k - \alpha_k A p_k \quad \hat{r}_{k+1} = \hat{r}_k - \alpha_k A^T \hat{p}_k$
$\beta_k = (\hat{r}_{k+1}, r_{k+1})/(\hat{r}_k, r_k)$
$p_{k+1} = r_{k+1} + \beta_k p_k, \quad \hat{p}_{k+1} = \hat{r}_{k+1} + \beta_k \hat{p}_k$.
end for

From above we have $(\tilde{A}\tilde{p}_k, \tilde{A}\tilde{p}_k)_o = (Ap_k, A^T p_k) \begin{pmatrix} 0 & A^{-T} \\ A^{-1} & 0 \end{pmatrix} \begin{pmatrix} Ap_k \\ A^T \hat{p}_k \end{pmatrix} = 2(\hat{p}_k, Ap_k)$.

BCG method satisfies the following relations:

$$r_k^T \hat{p}_i = \hat{r}_k^T p_i = 0, \quad i < k \tag{4.10.11a}$$
$$p_k^T A^T \hat{p}_i = \hat{p}_k^T A p_i = 0, \quad i < k \tag{4.10.11b}$$
$$r_k^T \hat{r}_i = \hat{r}_k^T r_i = 0, \quad i < k \tag{4.10.11c}$$

Definition 4.10.1 (4.10.11c) *and* (4.10.11b) *are called biorthogonality and biconjugacy condition, respectively.*

Property 4.10.1 **(i)** *In BCG method, the residual of the linear equation does not satisfy the minimal property, because P is taken by*

$$P = \tilde{A}^{-T} Z = \begin{pmatrix} 0 & A^{-T} \\ A^{-1} & 0 \end{pmatrix}$$

and P is symmetric, but not positive definite. The minimal value of the functional $f(x)$ may not exist.

(ii) *BCG method can break down, because $Z = (\tilde{A}^T P + P\tilde{A})/2$ is not positive definite. From above discussion, α_k can be zero. But this case occurs very few.*

GCG	
GCR, GCR(k)	BCG
Orthomin(k)	CGS
Orthodir	BiCGSTAB
Orthores	QMR
GMRES(m)	TFQMR
FOM	
Axelsson LS	

4.11 CGS (Conjugate Gradient Squared), A fast Lanczos-type solver for nonsymmetric linear systems

4.11.1 The polynomial equivalent method of the CG method

Consider first A is s.p.d. Then the CG method

$r_0 = b - Ax_0 = p_0$
for $i = 0, 1, 2, \cdots$
  $a_i = (r_i, p_i)/(p_i, Ap_i) = (r_i, r_i)/(p_i, Ap_i)$
  $x_{i+1} = x_i + a_i p_i$
  $r_{i+1} = r_i - a_i A p_i$
  $p_{i+1} = r_{i+1} + b_i p_i$
  $b_i = -(r_{i+1}, Ap_i)/(p_i, Ap_i) = -(r_{i+1}, r_{i+1})/(r_i, r_i)$

is equivalent to

$$
\begin{aligned}
&r_0 = b - Ax_0,\ p_{-1} = 1, \rho_{-1} = -1\\
&\text{for}\ \ n = 0, 1, 2, \cdots\\
&\quad \rho_n = r_n^T r_n,\ \ \beta_n = \rho_n/\rho_{n-1}\\
&\quad p_n = r_n + \beta_n p_{n-1}\\
&\quad \sigma_n = p_n^T A p_n,\ \ \alpha_n = \rho_n/\sigma_n\\
&\quad r_{n+1} = r_n - \alpha_n A p_n\\
&\quad x_{n+1} = x_n + \alpha_n p_n\ \ (r_n = b - Ax_n)
\end{aligned}
$$

Remark 4.11.1 **1.** $E_n = r_n^T A^{-1} r_n = \min_{x \in x_0 + K_n} \|b - Ax\|_{A^{-1}}$

2. $r_n^T r_m = \rho_n \delta_{nm}, \quad p_n^T A p_m = \sigma_n \delta_{nm}$

From the structure of the new form of the CG method, we write

$$r_n = \varphi_n(A) r_0, \qquad p_n = \psi_n(A) r_0$$

where φ_n and ψ_n are polynomial of degree $\leq n$. Define $\varphi_0(\tau) \equiv 1$ and $\varphi_{-1}(\tau) \equiv 0$. Then we find

$$p_n = \varphi_n(A) r_0 + \beta_n \psi_{n-1}(A) r_0 \equiv \psi_n(A) r_0 \tag{4.11.12a}$$

with

$$\psi_n(\tau) \equiv \varphi_n(\tau) + \beta_n \psi_{n-1}(\tau), \tag{4.11.12b}$$

and

$$r_{n+1} = \varphi_n(A) r_0 - \alpha_n A \psi_n(A) r_0 \equiv \varphi_{n+1}(A) r_0 \tag{4.11.13a}$$

with

$$\varphi_{n+1}(\tau) \equiv \varphi_n(\tau) - \alpha_n \tau \psi_n(\tau). \tag{4.11.13b}$$

The CG method can be re-iterpreted as an algorithm for generating a system of (orthogonal) polynomials. Define the symmetric bilinear form $(\cdot, \cdot)$ by

$$(\varphi, \psi) = [\varphi(A) r_0]^T \psi(A) r_0.$$

We have $(\varphi, \varphi) \geq 0$. Since A is symmetric, we can write

$$(\varphi, \psi) = r_0^T \varphi(A) \psi(A) r_0.$$

Furthermore, from the associate law of matrices

$$(\varphi\theta, \psi) = (\varphi, \theta\psi)$$

for any polynomial $\varphi,\ \theta\ , \psi$. Here $(\cdot, \cdot)$ is semidefinite, thus $(\varphi, \varphi) = 0$ may occur!

The polynomial equivalent method of the CG method :

$$
\begin{aligned}
&\varphi_0 \equiv 1,\ \varphi_{-1} \equiv 0,\ \rho_{-1} = 1\\
&\text{for}\ \ n = 0, 1, 2, \cdots\\
&\quad \rho_n = (\varphi_n, \varphi_n),\ \beta_n = \rho_n/\rho_{n-1}\\
&\quad \psi_n = \varphi_n + \beta_n \psi_{n-1}\\
&\quad \sigma_n = (\psi_n, \theta\psi_n),\ \alpha_n = \rho_n/\sigma_n\\
&\quad \varphi_{n+1} = \varphi_n - \alpha_n \theta \psi_n.
\end{aligned}
$$

where $\theta(\tau) = \tau$.

The minimization property reads

$$E_n = (\varphi_n, \theta^{-1}\varphi_n) = \min_{\varphi \in P^N} \frac{(\varphi, \theta^{-1}\varphi)}{\varphi(0)^2}.$$

We also have

$$(\varphi_i, \varphi_j) = 0, \quad i \neq j \quad \text{from} \quad (r_i, r_j) = 0, \quad i \neq j.$$
$$(\psi_i, \theta\psi_j) = 0, \quad i \neq j \quad \text{from} \quad (p_i, Ap_j) = 0, \quad i \neq j.$$

Theorem 4.11.1 *Let $[\cdot, \cdot]$ be any symmetric bilinear form satisfying*

$$[\varphi\chi, \psi] = [\varphi, \chi\psi] \quad \forall \varphi,\ \psi,\ \chi \in P^N$$

Let the sequence of φ_n and ψ_n be constructed according to PE algorithm, but using $[\cdot, \cdot]$ instead $(\cdot, \cdot)$. Then as long as the algorithm does not break down by zero division, then φ_n and ψ_n satisfy

$$[\varphi_n, \varphi_m] = \rho_n \delta_{nm}, \qquad [\psi_n, \theta\psi_m] = \sigma_n \delta_{nm}$$

with $\theta(\tau) \equiv \tau$.

Proof: By induction we prove the following statement:

$$[\psi_{n-1}, \theta\psi_k] = \sigma_{n-1}\delta_{n-1,k}, \quad [\varphi_n, \psi_k] = 0 \tag{4.11.14}$$

$\forall n \geq 0,\ -1 \leq k \leq n-1$ with $\sigma_{-1} = 0$. If $n = 0$, this is true since $\psi_{-1}(\tau) \equiv 0$. Suppose (4.11.14) holds for $n \leq m$ and let $k < m$. Then by PE algorithm, it holds

$$[\psi_m, \theta\psi_k] = [\varphi_m, \theta\psi_k] + \beta_m [\varphi_{m-1}, \theta\psi_k]. \tag{4.11.15}$$

Substitute $\theta\psi_k = (\varphi_k - \varphi_{k+1})/\alpha_k$ in the first term. The second term is zero for $k < m-1$, by hypothesis. Thus

$$[\psi_m, \theta\psi_k] = \frac{[\varphi_m, \varphi_k] - [\varphi_m, \varphi_{m+1}]}{\alpha_k} + \beta_m \sigma_{m-1}\delta_{m-1,k}, \quad \forall k \leq m-1.$$

If $k < m-1$, then $[\psi_m, \theta\psi_k] = 0$. For $k = m-1$ we have

$$[\psi_m, \theta\psi_{m-1}] = -\rho_m/\alpha_{m-1} + \beta_m \sigma_{m-1} = 0,$$

which proves first part of (4.11.14) for $n = m+1$.

Second Part: Write

$$[\varphi_{m+1}, \psi_k] = [\varphi_m, \psi_k] - \alpha_m [\psi_m, \theta\psi_k], \quad \forall k \leq m.$$

If $k \leq m-1$, then $[\varphi_{m+1}, \psi_k] = 0$ by hypothesis. Using the algorithm and choosing $k = m$ we get

$$[\varphi_{m+1}, \psi_m] = [\varphi_m, \varphi_m + \beta_m \psi_{m-1}] - \alpha_m [\psi_m, \theta\psi_m] = \rho_m - \alpha_m \sigma_m = 0,$$

which proves the second part of (4.11.14).

By Induction (4.11.14) is valid for all n. Finally, writing $\varphi_k = \psi_k - \beta_k \psi_{k-1}, \quad \forall k \geq 0$, it implies

$$[\varphi_n, \varphi_k] = [\varphi_n, \psi_k] - \beta_n [\varphi_n, \psi_{k-1}] = 0, \quad \forall k < n.$$

Together with the first part of (4.11.14), we prove the theorem. ■

The theorem is valid as long as the algorithm does not break down. For this reason we shall use orthogonal polynomial for φ_n and ψ_n, whether or not the bilinear forms involved are inner products.

In the following, we want to generalize the CG Algorithm to the nonsymmetric case. Consider

$$Ax = b, \quad A : \text{ nonsymmetric.}$$

Given $x_0,\ r_0 = b - Ax_0$, let $\tilde{r}_0$ be a suitably chosen vector. Define $[\cdot, \cdot]$ by

$$[\varphi, \psi] = \tilde{r}_0^T \varphi(A)\psi(A) r_0 = (\varphi(A^T)\tilde{r}_0)^T \psi(A) r_0$$

and define $p_{-1} = \tilde{p}_{-1} = 0$. (If A symmetric : $(\varphi, \psi) = r_0^T \varphi(A)\psi(A) r_0$). Then we have

$$\begin{aligned} r_n &= \varphi_n(A) r_0, \quad \tilde{r}_n = \varphi_n(A^T)\tilde{r}_0, \\ p_n &= \psi_n(A) r_0, \quad \tilde{p}_n = \psi_n(A^T)\tilde{r}_0 \end{aligned}$$

with φ_n and ψ_n according to (4.11.12b) and (4.11.13b). Indeed, these vectors can be produced by the Bi-Conjugate Gradient algorithm:

Algorithm 4.11.1 (Bi-Conjugate Gradient algorithm)

Given $r_0 = b - Ax_0$, $p_{-1} = \tilde{p}_{-1}$ and $\tilde{r}_0$ arbitrary
For $n = 0, 1, \cdots$
$\rho_n = \tilde{r}_n^T r_n, \quad \beta_n = \rho_n / \rho_{n-1}$
$p_n = r_n + \beta_n p_{n-1}, \quad \tilde{p}_n = \tilde{r}_n + \beta_n \tilde{p}_{n-1}$
$\sigma_n = \tilde{p}_n^T A p_n, \quad \alpha_n = \rho_n / \sigma_n$
$r_{n+1} = r_n - \alpha_n A p_n, \quad \tilde{r}_{n+1} = \tilde{r}_n - \alpha_n A^T \tilde{p}_n$
$x_{n+1} = x_n + \alpha_n p_n.$

Property 4.11.1 *$r_n = b - Ax_n$, $r_k^T \tilde{r}_j = 0,\ j \neq k$ and $p_k^T A^T \tilde{p}_j = 0,\ j \neq k$.*

Remark 4.11.2 *The Bi-Conjugate Gradient method is equivalent to the Lanczos biorthogonalization method.*

$$\begin{aligned} K_m &= span(V_m) = span(r_0, Ar_0, \cdots, A^{m-1} r_0) = span(p_0, p_1, \cdots, p_{m-1}), \\ L_m &= span(W_m) = span(\tilde{r}_0, A^T \tilde{r}_0, \cdots, (A^T)^{m-1} \tilde{r}_0) = span(\tilde{p}_0, \tilde{p}_1, \cdots, \tilde{p}_{m-1}). \end{aligned}$$

Remark 4.11.3 *In practice $\tilde{r}_0$ is often chosen equal to r_0. Then, if A is not too far from being S.P.D., the bilinear expressions $[\cdot, \cdot]$ and $[\cdot, \theta\cdot]$ will be positive semi-definite, and the algorithm will converge in the same way, and by the same argument as does the ordinary CG algorithm in the SPD-case!*

4.11.2 Squaring the CG algorithm: CGS Algorithm

Assume that Bi-CG is converging well. Then $r_n \to 0$ as $n \to \infty$. Because $r_n = \varphi_n(A)r_0$, $\varphi_n(A)$ behaves like contracting operators.

• Expect: $\varphi_n(A^T)$ behaves like contracting operators (i.e., $\tilde{r}_n \to 0$). But "quasi-residuals" $\tilde{r}_n$ is not exploited, they need to be computed for the ρ_n and σ_n.

• Disadvantage: Work of Bi-CG is twice the work of CG and in general $A^T v$ is not easy to compute. Especially if A is stored with a general data structure.

• Improvement: Using Polynomial equivalent algorithm to CG.

Since $\rho_n = [\varphi_n, \varphi_n]$ and $\sigma_n = [\psi_n, \theta\psi_n]$, $[\cdot,\cdot]$ has the property $[\varphi\chi, \psi] = [\varphi, \chi\psi]$. Let $\varphi_0 = 1$. Then

$$\rho_n = [\varphi_0, \varphi_n^2], \quad \sigma_n = [\varphi_0, \theta\psi_n^2].$$

$$\begin{cases} \varphi_{n+1} = \varphi_n - \alpha_n\theta\psi_n, \\ \psi_n = \varphi_n + \beta_n\psi_{n-1}. \end{cases}$$

• Purpose: (i) Find an algorithm that generates the polynomials φ_n^2 and ψ_n^2 rather than φ_n and ψ_n.
(ii) Compute the approximation solution x_n with $r_n = \varphi_n^2(A)r_0$ as residuals (try to interpret). Because $\rho_n = \tilde{r}_0^T r_n$ with $r_n = \varphi_n^2(A)r_0$, $\tilde{r}_n$ and $\tilde{p}_n$ need not to be computed. How to compute φ_n^2 and ψ_n^2?

$$\begin{aligned} \psi_n^2 &= [\varphi_n + \beta_n\psi_{n-1}]^2 = \varphi_n^2 + 2\beta_n\varphi_n\psi_{n-1} + \beta_n^2\psi_{n-1}^2, \\ \varphi_{n+1}^2 &= [\varphi_n - \alpha_n\theta\psi_n]^2 = \varphi_n^2 - 2\alpha_n\theta\varphi_n\psi_n + \alpha_n^2\theta^2\psi_n^2. \end{aligned}$$

Since

$$\varphi_n\psi_n = \varphi_n[\varphi_n + \beta_n\psi_{n-1}] = \varphi_n^2 + \beta_n\varphi_n\psi_{n-1},$$

we only need to compute $\varphi_n\psi_{n-1}$, φ_n^2 and ψ_n^2. Now define for $n \geq 0$:

$$\Phi_n = \varphi_n^2, \quad \Theta_n = \varphi_n\psi_{n-1}, \quad \Psi_{n-1} = \psi_{n-1}^2.$$

Algorithm 4.11.2 (CGS)

$\Phi_0 \equiv 1$. $\Theta_0 \equiv \Psi_{-1} \equiv 0, \rho_{-1} = 1$.
for $n = 0, 1, \cdots$
$\rho_n = [1, \Phi_n], \quad \beta_n = \rho_n/\rho_{n-1}$
$Y_n = \Phi_n + \beta_n\Theta_n$
$\Psi_n = Y_n + \beta_n(\Theta_n + \beta_n\Psi_{n-1})$
$\sigma_n = [1, \theta\Psi_n], \quad \alpha_n = \rho_n/\sigma_n, \quad \Theta(\tau) = \tau,$
$\Theta_{n+1} = Y_n - \alpha_n\theta\Psi_n$
$\Phi_{n+1} = \Phi_n - \alpha_n\theta(Y_n + \Theta_{n+1})$

Algorithm 4.11.3 (CGS Variant)

Define $r_n = \Phi_n(A)r_0, q_n = \Theta_n(A)r_0, p_n = \Psi_n(A)r_0,$
$r_0 = b - Ax_0, q_0 = p_{-1} = 0, \rho_{-1} = 1$
for $n = 0, 1, \cdots$
$\rho_n = \tilde{r}_0^T r_n, \quad \beta_n = \rho_n/\rho_{n-1}$
$u_n = r_n + \beta_n q_n$
$p_n = u_n + \beta_n(q_n + \beta_n p_{n-1})$
$v_n = Ap_n$
$\sigma_n = \tilde{r}_0^T v_n, \quad \alpha_n = \rho_n/\sigma_n$
$q_{n+1} = u_n - \alpha_n v_n$
$r_{n+1} = r_n - \alpha_n A(u_n + q_{n+1})$
$x_{n+1} = x_n + \alpha_n(u_n + q_{n+1}).$

Since $r_0 = b - Ax_0$, $r_{n+1} - r_n = A(x_n - x_{n+1})$, we have that $r_n = b - Ax_n$. So this algorithm produces x_n of which the residual satisfy

$$r_n = \varphi_n^2(A)r_0.$$

Remark 4.11.4 *Each step requires twice the amount of work necessary for symmetric CG. However the contracting effect of $\varphi_n(A)$ is used twice each step. The work is not more than for Bi-CG and working with A^T is avoided.*

4.12 Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems

Algorithm 4.12.1 (Bi-CG method)

Given $x_0,\ r_0 = b - Ax_0,\ (\tilde{r}_0, r_0) \neq 0,\ \rho_0 = 1,\ \tilde{p}_0 = p_0 = 0.$
For $i = 1, 2, 3, \cdots$
$\rho_i = (\tilde{r}_{i-1}, r_{i-1})$
$\beta_i = (\rho_i/\rho_{i-1})$
$p_i = r_{i-1} + \beta_i p_{i-1}$
$\tilde{p}_i = \tilde{r}_{i-1} + \beta_i \tilde{p}_{i-1}$
$v_i = Ap_i$
$\alpha_i = \rho_i/(\tilde{p}_i, v_i)$
$x_i = x_{i-1} + \alpha_i p_i$
Stop here, if x_i is accurate enough.
$r_i = r_{i-1} - \alpha_i v_i = r_{i-1} - \alpha_i A p_i$
$\tilde{r}_i = \tilde{r}_{i-1} - \alpha_i A^T \tilde{p}_i$
end for

Property 4.12.1 **(i)** $r_j \perp \tilde{r}_0, \ldots, \tilde{r}_{j-1}$ *and* $\tilde{r}_j \perp r_0, \ldots, r_{j-1}$.

(ii) *three-term recurrence relations between* $\{r_j\}$ *and* $\{\tilde{r}_j\}$.

(iii) *It terminates within n steps, but no minimal property.*

Since $r_j^{Bi-CG} = \varphi_j(A)r_0$ and $\tilde{r}_j^{Bi-CG} = \varphi_j(A^T)\tilde{r}_0$, it implies that

$$(r_j, \tilde{r}_i) = \big(\varphi_j(A)r_0, \varphi_i(A^T)\tilde{r}_0\big) = (\varphi_i(A)\varphi_j(A)r_0, \tilde{r}_0) = 0, \quad i < j.$$

Algorithm 4.12.2 (CGS method)

Given $x_0,\ r_0 = b - Ax_0,\ (r_0, \tilde{r}_0) \neq 0,\ \tilde{r}_0 = r_0,\ \rho_0 = 1,\ p_0 = q_0 = 0.$
For $i = 1, 2, 3, \cdots$
$\rho_i = (\tilde{r}_0, r_{i-1})$
$\beta = \rho_i/\rho_{i-1}$
$u = r_{i-1} + \beta q_{i-1}$
$p_i = u + \beta(q_{i-1} + \beta p_{i-1})$
$v = Ap_i$
$\alpha = \rho_i/(\tilde{r}_0, v)$
$q_i = u - \alpha v$
$w = u + q_i$
$x_i = x_{i-1} + \alpha w$
Stop here, if x_i *is accurate enough.*
$r_i = r_{i-1} - \alpha A w$
end for

We have $r_j^{\mathrm{CGS}} = \varphi_i(A)^2 r_0$.
From Bi-CG method we have $r_i^{Bi-CG} = \varphi_i(A)r_0$ and $p_{i+1} = \psi_i(A)r_0$. Thus we get

$$\psi_i(A)r_0 = \left(\varphi_i(A) + \beta_{i+1}\psi_{i-1}(A)\right) r_0,$$

and

$$\varphi_i(A)r_0 = \left(\varphi_{i-1}(A) - \alpha_i A\psi_{i-1}(A)\right) r_0,$$

where $\psi_i = \varphi_i + \beta_{i+1}\psi_{i-1}$ and $\varphi_i = \varphi_{i-1} - \alpha_i\theta\psi_{i-1}$. Since

$$\big(\varphi_i(A)r_0, \varphi_j(A^T)\tilde{r}_0\big) = 0, \quad j < i,$$

it holds that

$$\varphi_i(A)r_0 \perp \tilde{r}_0, A^T\tilde{r}_0, \ldots, (A^T)^{i-1}\tilde{r}_0$$

if and only if

$$(\tilde{\varphi}_j(A)\varphi_i(A)r_0, \tilde{r}_0,) = 0$$

for some polynomial $\tilde{\varphi}_j$ of degree $j < i$ for $j = 0, 1, \cdots, i-1$. In Bi-CG method, we take $\tilde{\varphi}_j = \varphi_j\ \tilde{r}_j = \varphi_j(A^T)\tilde{r}_0$ and exploit it in CGS to get $r_j^{CGS} = \varphi_j^2(A)r_0$. Now $r_i = \tilde{\varphi}_i(A)\varphi_i(A)r_0$. How to choose $\tilde{\varphi}_i$ polynomial of degree i so that $\|r_i\|$ satisfies the minimum. Like polynomial, we can determine the optimal parameters of $\tilde{\varphi}_i$ so that $\|r_i\|$ satisfies the minimum. But the optimal parameters for the Chebychev polynomial are in general not easily obtainable. Now we take

$$\tilde{\varphi}_i \equiv \eta_i(x),$$

where

$\eta_i(x) = (1-\omega_1 x)(1-\omega_2 x)\cdots(1-\omega_i x)$.

Here ω_j are suitable constants to be selected.

Define

$$r_j = \eta_j(A)\varphi_j(A)r_0.$$

Then

$$\begin{aligned} r_i &= \eta_i(A)\varphi_i(A)r_0 \\ &= (1-\omega_i A)\eta_{i-1}(A)\left(\varphi_{i-1}(A) - \alpha_i A\psi_{i-1}(A)\right)r_0 \\ &= \left\{(\eta_{i-1}(A)\varphi_{i-1}(A) - \alpha_i A\eta_{i-1}(A)\psi_{i-1}(A))\right\}r_0 \\ &\quad -\omega_i A\left\{(\eta_{i-1}(A)\varphi_{i-1}(A) - \alpha_i A\eta_{i-1}(A)\psi_{i-1}(A))\right\}r_0 \\ &= r_{i-1} - \alpha_i A p_i - \omega_i A(r_{i-1} - \alpha_i A p_i) \end{aligned}$$

and

$$\begin{aligned} p_{i+1} &= \eta_i(A)\psi_i(A)r_0 \\ &= \eta_i(A)\left(\varphi_i(A) + \beta_{i+1}\psi_{i-1}(A)\right)r_0 \\ &= \eta_i(A)\varphi_i(A)r_0 + \beta_{i+1}(1-\omega_i A)\eta_{i-1}(A)\psi_{i-1}(A)r_0 \\ &= \eta_i(A)\varphi_i(A)r_0 + \beta_{i+1}\eta_{i-1}(A)\psi_{i-1}(A)r_0 \\ &\quad -\beta_{i+1}\omega_i A\eta_{i-1}(A)\psi_{i-1}(A)r_0 \\ &= r_i + \beta_{i+1}(p_i - \omega_i A p_i). \end{aligned}$$

Recover the constants ρ_i, β_i, and α_i in Bi-CG method. We now compute β_i: Let

$$\tilde{\rho}_{i+1} = (\tilde{r}_0, \eta_i(A)\varphi_i(A)r_0) = \left(\eta_i(A^T)\tilde{r}_0, \varphi_i(A)r_0\right).$$

From Bi-CG we have $\varphi_i(A)r_0 \perp$ all vectors $\mu_{i-1}(A^T)\tilde{r}_0$, where μ_{i-1} is an arbitrary polynomial of degree $i-1$. Consider the highest order term of $\eta_i(A^T)$ (when computing $\tilde{\rho}_{i+1}$) is $(-1)^i\omega_1\omega_2\cdots\omega_i(A^T)^i$. From Bi-CG method, we also have

$$\rho_{i+1} = \left(\varphi_i(A^T)\tilde{r}_0, \varphi_i(A)r_0\right).$$

The highest order term of $\varphi_i(A^T)$ is $(-1)^i\alpha_1\cdots\alpha_i(A^T)^i$. Thus

$$\beta_i = (\tilde{\rho}_i/\tilde{\rho}_{i-1})\,(\alpha_{i-1}/\omega_{i-1}),$$

because

$$\begin{aligned} \beta_i &= \frac{\rho_i}{\rho_{i-1}} = \frac{\left(\alpha_1\cdots\alpha_{i-1}(A^T)^{i-1}\tilde{r}_0, \varphi_{i-1}(A)r_0\right)}{\left(\alpha_1\cdots\alpha_{i-2}(A^T)^{i-2}\tilde{r}_0, \varphi_{i-2}(A)r_0\right)} \\ &= \frac{\left(\frac{\alpha_1\cdots\alpha_{i-1}}{\omega_1\cdots\omega_{i-1}}\omega_1\cdots\omega_{i-1}(A^T)^{i-1}\tilde{r}_0, \varphi_{i-1}(A)r_0\right)}{\left(\frac{\alpha_1\cdots\alpha_{i-2}}{\omega_1\cdots\omega_{i-2}}\omega_1\cdots\omega_{i-2}(A^T)^{i-2}\tilde{r}_0, \varphi_{i-2}(A)r_0\right)} \\ &= (\tilde{\rho}_i/\tilde{\rho}_{i-1})\,(\alpha_{i-1}/\omega_{i-1}). \end{aligned}$$

Similarly, we can compute ρ_i and α_i. Let

$$r_i = r_{i-1} - \gamma A y, \quad x_i = x_{i-1} + \gamma y \quad \text{(side product)}.$$

Compute ω_i so that $r_i = \eta_i(A)\varphi(A)r_0$ is minimized in 2-norm as a function of ω_i.

Algorithm 4.12.3 (Bi-CGSTAB method)

Given $x_0,\ r_0 = b - Ax_0,\ \tilde{r}_0$ *arbitrary, such that* $(\tilde{r}_0, r_0) \neq 0,$ *e.g.* $\tilde{r}_0 = r_0,$
$\rho_0 = \alpha = \omega_0 = 1,\ v_0 = p_0 = 0$
For $i = 1, 2, 3, \cdots$
$\rho_i = (\tilde{r}_0, r_{i-1})$
$\beta = (\rho_i/\rho_{i-1})(\alpha/\omega_{i-1})$
$p_i = r_{i-1} + \beta(p_{i-1} - \omega_{i-1}v_{i-1})$
$v_i = Ap_i$
$\alpha = \rho_i/(\tilde{r}_0, v_i)$
$s = r_{i-1} - \alpha v_i$
$t = As$
$\omega_i = (t, s)/(t, t)$
$x_i = x_{i-1} + \alpha p_i + \omega_i s \quad (= x_{i-1} + \alpha p_i + \omega_i(r_{i-1} - \alpha A p_i))$
Stop here, if x_i *is accurate enough.*
$r_i = s - \omega_i t\ [= r_{i-1} - \alpha A p_i - \omega_i A(r_{i-1} - \alpha A p_i) = r_{i-1} - A(\alpha p_i + \omega_i(r_{i-1} - \alpha A p_i)]$
end for

Preconditioned Bi-CGSTAB-P:
Rewrite $Ax = b$ as

$$\tilde{A}\tilde{x} = \tilde{b} \quad \text{with} \quad \tilde{A} = K_1^{-1} A K_2^{-1},$$

where $x = K_2^{-1}\tilde{x}$ and $\tilde{b} = K_1^{-1}b$. Then

$$\begin{aligned} &\tilde{p}_i \Rightarrow K_1^{-1}p_i, \quad \tilde{v}_i \Rightarrow K_1^{-1}v_i, \quad \tilde{r}_i \Rightarrow K_1^{-1}r_i, \\ &\tilde{s} \Rightarrow K_1^{-1}s_i, \quad \tilde{t} \Rightarrow K_1^{-1}t_i, \quad \tilde{x} \Rightarrow K_2 x_i, \\ &\tilde{r}_0 \Rightarrow K_1^T \hat{r}_0. \end{aligned}$$

4.13 A Transpose-Free Qusi-minimal Residual Algorithm for Nonsymmetric Linear Systems

Given x_0, $r_0 = b - Ax_0$ and $\tilde{r}_0$ arbitrary such that $\tilde{r}_0^T r_0 \neq 0$, e.g. $\tilde{r}_0 = r_0$. We know that

$$w^T(b - Ax_n^{BCG}) = 0, \quad \forall\ w \in K_n(\tilde{r}_0, A^T), \quad x_n^{BCG} \in x_0 + K_n(r_0, A).$$

The nth iterate, x_n^{BCG}, generated by Bi-CG is defined by Petrov-Galerkin method.

$$\begin{aligned} r_n^{BCG} &= \varphi_n(A)r_0, \quad \varphi_n \in P_n, \quad \varphi_n(0) = 1. \\ r_n^{CGS} &= (\varphi_n(A))^2 r_0, \quad x_n \in x_0 + K_{2n}(r_0, A). \\ r_n^{BiCGSTAB} &= \eta_n(A)\varphi_n(A)r_0, \quad x_n \in x_0 + K_{2n}(r_0, A). \end{aligned}$$

Algorithm 4.13.1 (CGS Algorithm)

Choose $x_0 \in \mathbb{R}^N$, *set* $p_0 = u_0 = r_0 = b - Ax_0$, $v_0 = Ap_0$,
Choose $\tilde{r}_0$ *such that* $\rho_0 = \tilde{r}^T r_0 \neq 0$,
for $n = 0, 1, 2, \cdots$
$\sigma_{n-1} = \tilde{r}_0^T v_{n-1}$, $\alpha_{n-1} = \rho_{n-1}/\sigma_{n-1}$,
$q_n = u_{n-1} - \alpha_{n-1} v_{n-1}$,
$x_n = x_{n-1} + \alpha_{n-1}(u_{n-1} + q_n)$,
$r_n = r_{n-1} - \alpha_{n-1} A(u_{n-1} + q_n)$,
If x_n *converges, stop;*
$\rho_n = \tilde{r}_0^T r_n$, $\beta_n = \rho_n/\rho_{n-1}$,
$u_n = r_n + \beta_n q_n$,
$p_n = u_n + \beta_n(q_n + \beta_n p_{n-1})$,
$v_n = Ap_n$.
end for

Note that

$$\alpha_{n-1} \neq 0 \quad \text{for all } n, \tag{4.13.1}$$

and

$$u_{n-1} = \varphi_{n-1}(A)\psi_{n-1}(A)r_0, \quad q_n = \varphi_n(A)\psi_{n-1}(A)r_0, \tag{4.13.2}$$

where φ_n, ψ_n are generated by

$$\psi_n(\tau) = \varphi_n(\tau) + \beta_n \psi_{n-1}(\tau), \quad \psi_0 \equiv 1 \tag{4.13.3}$$

and

$$\varphi_n(\tau) = \varphi_{n-1}(\tau) - \alpha_{n-1}\tau\psi_{n-1}(\tau). \tag{4.13.4}$$

4.13.1 Quasi-Minimal Residual Approach

Set

$$y_m = \begin{cases} u_{n-1}, & \text{if } m = 2n-1, \text{ odd} \\ q_n, & \text{if } m = 2n, \text{ even} \end{cases} \tag{4.13.5}$$

and

$$w_m = \begin{cases} \varphi_n^2(A)r_0, & \text{if } m = 2n+1, \text{ odd} \\ \varphi_n(A)\varphi_{n-1}(A)r_0, & \text{if } m = 2n, \text{ even} \end{cases} \tag{4.13.6}$$

From $r_n^{CGS} = \varphi_n^2(A)r_0$ follows that $w_{2n+1} = r_n^{CGS}$. Using (4.13.2) and (4.13.4) we get

$$\psi_{n-1}(A) = A^{-1}\frac{1}{\alpha_{n-1}}(\varphi_{n-1}(A) - \varphi_n(A)).$$

Multiply above equation by $\varphi_n(A)$, then the vectors in (4.13.5) and (4.13.6) are related by

$$Ay_m = \frac{1}{\alpha_{\lfloor (m-1)/2 \rfloor}}(w_m - w_{m+1}). \tag{4.13.7}$$

By (4.13.1), $\alpha_{\lfloor (m-1)/2 \rfloor}$ in (4.13.7) $\neq 0$. Let

$$Y_m = [y_1, y_2, \cdots, y_m], \quad W_{m+1} = [w_1, \cdots, w_m, w_{m+1}].$$

Then from (4.13.7) we get

$$AY_m = W_{m+1} B_m^{(e)}, \tag{4.13.8}$$

where

$$B_m^{(e)} = \begin{pmatrix} 1 & & & & 0 \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & & -1 & 1 \\ 0 & & & & -1 \end{pmatrix} \operatorname{diag}(\alpha_0, \alpha_0, \alpha_1, \alpha_1, \cdots, \alpha_{\lfloor (m-1)/2 \rfloor})^{-1} \tag{4.13.9}$$

is an $(m+1) \times m$ lower bidiagonal matrix.

By (4.13.3), (4.13.4) and (4.13.1) we have that polynomials φ_n and ψ_n are of full degree n. With (4.13.2) and (4.13.5) it implies

$$K_m(r_0, A) = span\{y_1, y_2, \cdots, y_m\} = \{Y_m z \mid z \in \mathbb{R}^m\}. \tag{4.13.10}$$

But any possible iterate x_m must lie in $x_0 + K_m(r_0, A)$. Thus

$$x_m = x_0 + Y_m z \quad \text{for some } z \in \mathbb{R}^m. \tag{4.13.11}$$

From (4.13.8) and $w_1 = r_0$ (see $w_{2n+1} = r_n^{CGS}$) follows that the residual satisfies

$$r_m = r_0 - AY_m z = W_{m+1}(e_1^{(m+1)} - B_m^{(e)} z). \tag{4.13.12}$$

Let

$$\Omega_{m+1} = \operatorname{diag}(w_1, w_2, \cdots, w_{m+1}), \quad w_k > 0, \tag{4.13.13}$$

be any scaling matrix, rewrite (4.13.12) as

$$r_m = W_{m+1} \Omega_{m+1}^{-1} (f_{m+1} - H_m^{(e)} z), \tag{4.13.14}$$

where

$$f_{m+1} = \omega_1 e_1^{(m+1)}, \quad H_m^{(e)} = \Omega_{m+1} B_m^{(e)}. \tag{4.13.15}$$

We now define the m-th iterate, x_m, of the transpose-free quasi-minimal residual method (TFQMR) by

$$x_m = x_0 + Y_m z_m, \tag{4.13.16}$$

where z_m is the solution of the least squares problem

$$\tau_m := \|f_{m+1} - H_m^{(e)} z_m\|_2 = \min_{z \in \mathbb{R}^m} \|f_{m+1} - H_m^{(e)} z\|_2 \tag{4.13.17}$$

By (4.13.9), (4.13.13) and (4.13.15) it implies that $H_m^{(e)}$ has full column rank m. Then z_m is uniquely defined by (4.13.17). In general, we set

$$w_k = \|w_k\|_2, \quad k = 1, \cdots, m+1.$$

This implies that all columns of $W_{m+1} \Omega_{m+1}^{-1}$ are unit vectors.

4.13.2 Derivation of actual implementation of TFQMR

Consider

$$\tilde{x}_m = x_0 + Y_m \tilde{z}_m, \quad \tilde{z}_m = H_m^{-1} f_m, \tag{4.13.18}$$

where

$$H_m^{(e)} = \begin{bmatrix} H_m \\ * \cdots * \end{bmatrix} \text{ and } f_{m+1} = \begin{bmatrix} f_m \\ * \end{bmatrix}.$$

By (4.13.9), (4.13.13) and (4.13.15) follows H_m nonsingular, thus

$$\tilde{z}_m = [\alpha_0, \alpha_0, \alpha_1, \cdots, \alpha_{\lfloor (m-1)/2 \rfloor}]^T \tag{4.13.19}$$

and

$$\omega_{m+1} = \|f_{m+1} - H_m^{(e)} \tilde{z}_m\|_2. \tag{4.13.20}$$

Comparing (4.13.18) and (4.13.19) with update formula for iterate x_n^{CGS} in CGS Algorithm we get

$$\tilde{x}_{2n} = x_n^{CGS}. \tag{4.13.21}$$

Lemma 4.13.1 *Let $w_1 > 0, m \geq 1$ and*

$$H_m^{(e)} = \begin{pmatrix} H_m & \\ h_{m+1,m} & e_m^T \end{pmatrix} = \begin{pmatrix} H_{m-1}^{(e)} & * \\ 0 & h_{m+1,m} \end{pmatrix} \tag{4.13.22}$$

be an $(m+1) \times m$ upper Hessenberg matrix of full column rank m. For $k = m-1, m$, let $z_k \in \mathbb{R}^k$ denote the solution of the least-square problem

$$\tau_k := \min_{z \in \mathbb{R}^k} \|f_{k+1} - H_k^{(e)} z\|_2, \quad f_{k+1} = w_1 e_1^{k+1} \in \mathbb{R}^{k+1}. \tag{4.13.23}$$

Moreover, assume that H_m in (4.13.22) is nonsingular. Set $\tilde{z}_m := H_m^{-1} f_m$. Then

$$z_m = (1 - c_m^2) \begin{pmatrix} z_{m-1} \\ 0 \end{pmatrix} + c_m^2 \tilde{z}_m^2, \tag{4.13.24}$$

$$\tau_m = \tau_{m-1} \theta_m c_m, \tag{4.13.25}$$

where

$$\theta_m = \frac{1}{\tau_{m-1}} \|f_{m+1} - H_m^{(e)} \tilde{z}_m\|_2, \quad c_m = \frac{1}{\sqrt{1 + \theta_m^2}}. \tag{4.13.26}$$

4.13.3 TFQMR Algorithm

From (4.13.24), (4.13.11) and (4.13.18) are connected by

$$x_m = (1 - c_m^2) x_{m-1} + c_m^2 \tilde{x}_m. \tag{4.13.27}$$

By (4.13.25), (4.13.26) and (4.13.20) follows that

$$\theta_m = \frac{w_{m+1}}{\tau_{m-1}}, \quad c_m = \frac{1}{\sqrt{1 + \theta_m^2}} \quad \text{and} \quad \tau_m = \tau_{m-1} \theta_m c_m. \tag{4.13.28}$$

Setting

$$d_m = \frac{1}{\alpha_{\lfloor (m-1)/2 \rfloor}}(\tilde{x}_m - x_{m-1}). \tag{4.13.29}$$

Rewrite (4.13.27) and get

$$x_m = x_{m-1} + \eta_m d_m, \tag{4.13.30}$$

where $\eta_m = c_m^2 \alpha_{\lfloor (m-1)/2 \rfloor}$. By (4.13.18) and (4.13.19) we get

$$\tilde{x}_m = x_0 + Y_m \tilde{z}_m, \quad \tilde{z}_m = [\alpha_0, \alpha_1, \cdots, \alpha_{\lfloor (m-1)/2 \rfloor}]^T,$$

and thus

$$\tilde{x}_m = \tilde{x}_{m-1} + \alpha_{\lfloor (m-1)/2 \rfloor} y_m.$$

Together with (4.13.29) and (4.13.30) (m replaced by $m-1$) we have

$$d_m = y_m + \frac{\theta_{m-1}^2 \eta_{m-1}}{\alpha_{\lfloor (m-1)/2 \rfloor}} d_{m-1}, \tag{4.13.31}$$

where $\theta_{m-1}^2 := \frac{1-c_m^2}{c_{m-1}^2}$.

Remark 4.13.1

$$\begin{aligned} d_m &= \frac{1}{\alpha}(\tilde{x}_{m-1} + \alpha y_m - x_{m-1}) = y_m + \frac{1}{\alpha}[\tilde{x}_{m-1} - x_{m-1}] \\ &= y_m + \frac{1}{\alpha}(\tilde{x}_{m-1} - x_{m-2} - \eta_{m-1} d_{m-1}) \\ &= y_m + \frac{1}{\alpha}(\tilde{\alpha} d_{m-1} - \eta_{m-1} d_{m-1}) = y_m + \frac{1}{\alpha}(\tilde{\alpha} - \eta_{m-1}) d_{m-1} \\ &= y_m + \frac{1}{\alpha}(\frac{\eta_{m-1}}{c_{m-1}^2} - \eta_{m-1}) d_{m-1} = y_m + \frac{1}{\alpha}(\eta_{m-1}(\frac{1-c_{m-1}^2}{c_{m-1}^2})) d_{m-1}. \end{aligned}$$

From (4.13.5) and (4.13.6), q_n and u_n in CGS Algorithm follows

$$y_{2n} = y_{2n-1} - \alpha_{n-1} v_{n-1}, \quad y_{2n+1} = w_{2n+1} + \beta_n v_{2n}. \tag{4.13.32}$$

Multiplying the update formula for p_n in CGS Algorithm by A we get

$$v_n = A y_{2n+1} + \beta_n (A y_{2n} + \beta_n v_{n-1}), \text{ for } v_n = A p_n. \tag{4.13.33}$$

By (4.13.7) w_m's can be generated by

$$w_{m+1} = w_m - \alpha_{\lfloor (m-1)/2 \rfloor} A y_m. \tag{4.13.34}$$

Combining (4.13.28), (4.13.30)-(4.13.34) we get the TFQMR Algorithm in standard weighting strategy $\omega_k = \|w_k\|_2$.

Algorithm 4.13.2 (TFQMR Algorithm)

Choose $x_0 \in \mathbb{R}^N$.
Set $w_1 = y_1 = r_0 = b - Ax_0,\ v_0 = Ay_1,\ d_0 = 0,\ \tau_0 = \|r_0\|_2,\ \theta_0 = 0,\ \eta_0 = 0;$
Choose $\tilde{r}_0$ *such that* $\rho_0 = \tilde{r}^T r_0 \neq 0,$
For $n = 0, 1, 2, \cdots$ *do*
$\quad$ *set* $\sigma_{n-1} = \tilde{r_0}^T v_{n-1},\ \alpha_{n-1} = \rho_{n-1}/\sigma_{n-1},\ y_{2n} = y_{2n-1} - \alpha_{n-1} v_{n-1},$
$\quad$ *For* $m = 2n-1, 2n$ *do*
$\quad\quad$ *set* $w_{m+1} = w_m - \alpha_{n-1} A y_m,$
$\quad\quad$ $\theta_m = \|w_{m+1}\|_2/\tau_{m-1},\ c_m = 1/\sqrt{1+\theta_m^2},$
$\quad\quad$ $\tau_m = \tau_{m-1}\theta_m c_m,\ \eta_m = c_m^2 \alpha_{n-1},$
$\quad\quad$ $d_m = y_m + (\theta_{m-1}^2 \eta_{m-1}/\alpha_{n-1}) d_{m-1},$
$\quad\quad$ $x_m = x_{m-1} + \eta_m d_m,$
$\quad\quad$ *If* x_m *converges, stop;*
$\quad$ *End for*
$\quad$ *set* $\rho_n = \tilde{r}_0^T w_{2n+1},\ \beta_n = \rho_n/\rho_{n-1},$
$\quad$ $y_{2n+1} = w_{2n+1} + \beta_n y_c 2n,$
$\quad$ $v_n = Ay_{2n+1} + \beta_n(Ay_{2n} + \beta_n v_{n-1}).$
End for

4.14 GMRES: Generalized Minimal Residual Algorithm for solving Nonsymmetric Linear Systems

Algorithm 4.14.1 (GCR)

Input: Given x_0, *compute* $p_0 = r_0 = b - Ax_0;$
Output: solution of linear system $Ax = b.$
$\quad$ *Iterate* $i = 0, 1, 2, \cdots,$
$\quad\quad$ *compute* $\alpha_i = (r_i, Ap_i)/(Ap_i, Ap_i),$
$\quad\quad\quad$ $x_{i+1} = x_i + \alpha_i p_i,$
$\quad\quad\quad$ $r_{i+1} = r_i - \alpha_i Ap_i \equiv b - Ax_i,$
$\quad\quad\quad$ $p_{i+1} = r_{i+1} + \sum_{j=0}^{i} \beta_j^{(i)} p_j,$
$\quad\quad$ $\beta_j^{(i)}$ *are chosen so that* $(Ap_{i+1}, Ap_j) = 0$ *, for* $0 \leq j \leq i.$
$\quad$ *End;*

It requires that $\frac{1}{2}(A^T + A)$ is symmetric positive definite.

Example 4.14.1 *Let*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Take $x_0 = 0$. *Then we obtain the following results:*

- *For* $i = 0$ *in Algorithm 4.14.1, we have that* $\alpha_0 = 0$ *which implies that* $x_1 = x_0$ *and* $r_1 = r_0$. *Thus* $p_1 = 0$.
- *For* $i = 1$ *in Algorithm 4.14.1, we see that a division by zero when computing* α_1 *and break down.*

4.14.1 FOM algorithm: Full orthogonalization method

For GMRES method,

(a) CANNOT break down, unless it has already converged.

(b) 1/2 storage required than GCR,

(c) 1/3 fewer arithmetic operations than GCR

Main goal: Find orthogonal basis for $K_k = \{r_0, Ar_0, \cdots, A^{k-1}r_0\}$, i.e., span$(K_k) =< v_1, \cdots, v_k >$, where $v_i \perp v_j$ for $i \neq j$.

Theorem 4.14.1 (Implicit Q theorem) *Let $AQ_1 = Q_1H_1$ and $AQ_2 = Q_2H_2$, where H_1, H_2 are Hessenberg and Q_1, Q_2 are unitary with $Q_1e_1 = Q_2e_1 = q_1$. Then $Q_1 = Q_2$ and $H_1 = H_2$.*

Proof: Let

$$A[q_1 \;\; q_2 \;\; \cdots \;\; q_n] = [q_1 \;\; q_2 \;\; \cdots \;\; q_n] \begin{bmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1n} \\ h_{21} & h_{22} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & h_{n-1,n} \\ 0 & \cdots & 0 & h_{n,n-1} & h_{nn} \end{bmatrix}. \tag{4.14.1}$$

Then we have

$$Aq_1 = h_{11}q_1 + h_{21}q_2. \tag{4.14.2}$$

Since $q_1 \perp q_2$, it implies that

$$h_{11} = q_1^* A q_1 / q_1^* q_1.$$

From (4.14.2), we get that

$$\tilde{q_2} \equiv h_{21}q_2 = Aq_1 - h_{11}q_1.$$

That is

$$q_2 = \tilde{q_2}/\|\tilde{q_2}\|_2 \quad \text{and} \quad h_{21} = \|\tilde{q_2}\|_2.$$

Similarly, from (4.14.1),

$$Aq_2 = h_{12}q_1 + h_{22}q_2 + h_{32}q_3,$$

where

$$h_{12} = q_1^* A q_2 \quad \text{and} \quad h_{22} = q_2^* A q_2.$$

Let

$$\tilde{q_3} = Aq_2 - h_{12}q_1 + h_{22}q_2.$$

Then

$$q_3 = \tilde{q_3}/\|\tilde{q_3}\|_2 \quad \text{and} \quad h_{32} = \|\tilde{q_3}\|,$$

and so on. Therefore, $[q_1, \cdots, q_n]$ are uniquely determined by q_1. Thus, uniqueness holds.

Let $K_n = [v_1, Av_1, \cdots, A^{n-1}v_1]$ with $\|v_1\|_2 = 1$ is nonsingular. $K_n = U_nR_n$ and $U_ne_1 = v_1$. Then

$$AK_n = K_nC_n = [v_1, Av_1, \cdots, A^{n-1}v_1] \begin{bmatrix} 0 & \cdots & \cdots & 0 & * \\ 1 & \ddots & & \vdots & * \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & * \end{bmatrix}. \tag{4.14.3}$$

Since K_n is nonsingular, (4.14.3) implies that

$$A = K_nC_nK_n^{-1} = (U_nR_n)C_n(R_n^{-1}U_n^{-1}).$$

That is

$$AU_n = U_n(R_nC_nR_n^{-1}),$$

where $(R_nC_nR_n^{-1})$ is Hessenberg and $U_ne_1 = v_1$. Because $< U_n >=< K_n >$, find $AV_n = V_nH_n$ by any method with $V_ne_1 = v_1$, then it holds that $V_n = U_n$, i.e., $v_n^{(i)} = u_n^{(i)}$ for $i = 1, \cdots, n$. ■

Algorithm 4.14.2 (Arnoldi algorithm)

Input: Given v_1 with $\|v_1\|_2 = 1$;
Output: Arnoldi factorization: $AV_k = V_kH_k + h_{k+1,k}v_{k+1}e_k^T$.
Iterate $j = 1, 2, \cdots,$
compute $h_{ij} = (Av_j, v_i)$ for $i = 1, 2, \cdots, j,$
$\tilde{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i,$
$h_{j+1,j} = \|\tilde{v}_{j+1}\|_2,$
$v_{j+1} = \tilde{v}_{j+1}/h_{j+1,j}.$
End;

Remark 4.14.1 (a) *Let $V_k = [v_1, \cdots, v_k] \in \mathbb{R}^{n\times k}$ where v_j, for $j = 1, \ldots, k$, is generated by Arnoldi algorithm. Then $H_k \equiv V_k^TAV_k$ is upper $k \times k$ Hessenberg.*

(b) *Arnoldi's original method was a Galerkin method for approximate the eigenvalue of A by H_k.*

In order to solve $Ax = b$ by the Galerkin method using $< K_k >\equiv< V_k >$, we seek an approximate solution $x_k = x_0 + z_k$ with $z_k \in K_k =< r_0, Ar_0, \cdots, A^{k-1}r_0 >$ and $r_0 = b - Ax_0$.

Definition 4.14.1 *$\{x_k\}$ is said to be satisfied the Galerkin condition if $r_k \equiv b - Ax_k$ is orthogonal to K_k for each k.*

The Galerkin method can be stated as that find

$$x_k = x_0 + z_k \quad \text{with} \quad z_k \in V_k \tag{4.14.4}$$

such that

$$(b - Ax_k, v) = 0, \quad \forall\ v \in V_k,$$

which is equivalent to find

$$z_k \equiv V_k y_k \in V_k \tag{4.14.5}$$

such that

$$(r_0 - Az_k, v) = 0, \quad \forall\ v \in V_k. \tag{4.14.6}$$

Substituting (4.14.5) into (4.14.6), we get

$$V_k^T(r_0 - AV_k y_k) = 0,$$

which implies that

$$y_k = (V_k^T A V_k)^{-1}\|r_0\|e_1. \tag{4.14.7}$$

Since V_k is computed by the Arnoldi algorithm with $v_1 = r_0/\|r_0\|$, y_k in (4.14.7) can be represented as

$$y_k = H_k^{-1}\|r_0\|e_1.$$

Substituting it into (4.14.5) and (4.14.4), we get

$$x_k = x_0 + V_k H_k^{-1}\|r_0\|e_1.$$

Using the result that $AV_k = V_k H_k + h_{k+1,k}v_{k+1}e_k^T$, r_k can be reformulated as

$$\begin{aligned} r_k &= b - Ax_k = r_0 - AV_k y_k = r_0 - (V_k H_k + h_{k+1,k}v_{k+1}e_k^T)y_k \\ &= r_0 - V_k\|r_0\|e_1 - h_{k+1,k}e_k^T y_k v_{k+1} = -(h_{k+1,k}e_k^T y_k)v_{k+1}. \end{aligned}$$

Algorithm 4.14.3 (FOM algorithm: Full orthogonalization method)

Input: *choose x_0, compute $r_0 = b - Ax_0$ and $v_1 = r_0/\|r_0\|$;*
Output: *solution of linear system $Ax = b$.*
  Iterate $j = 1, 2, \cdots, k$,
    compute $h_{ij} = (Av_j, v_i)$ *for* $i = 1, 2, \cdots, j$,
      $\tilde{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i$,
      $h_{j+1,j} = \|\tilde{v}_{j+1}\|_2$,
      $v_{j+1} = \tilde{v}_{j+1}/h_{j+1,j}$.
  End;
  Form the solution:
      $x_k = x_0 + V_k y_k$, *where* $y_k = \|r_0\|H_k^{-1}e_1$.

In practice, k is chosen such that the approximate solution x_k will be sufficiently accurate. Fortunately, it is simple to determine a posteriori when k is sufficiently large without having to explicitly compute x_k. Furthermore, we have

$$\|b - Ax_k\| = h_{k+1,k}|e_k^{\mathrm{T}} y_k|$$

Property 4.14.1 (FOM) **(a)** $r_k // v_{k+1} \Rightarrow r_i \perp r_j, \quad i \neq j$

(b) *FOM does NOT break down* $\iff$ *If the degree of the minimal polynomial of* v_1 *is at least* k*, and the matrix* H_k *is nonsingular.*

(c) *The process terminates at most* N *steps.*

A difficulty with the full orthogonalization method is that it becomes increasingly expensive when k increases. There are two distinct ways of avoiding this difficulty.

(i) restart the algorithm every m steps

(ii) v_{i+1} are only orthogonal to the previous ℓ vectors. H_k is then banded, then we have incomplete FOM(ℓ).

A drawback of these truncation techniques is the lack of any theory concerning the global convergence of these truncation technique. Such a theory is difficult because there is NO optimality property similar to that of CG method. Therefore, we consider GMRES which satisfies an optimality property.

4.14.2 The generalized minimal residual (GMRES) algorithm

The approximate solution of the form $x_0 + z_k$, which minimizes the residual norm over $z_k \in K_k$, can in principle be obtained by following algorithms:

- The ORTHODIR algorithm of Jea and Young;
- the generalized conjugate residual method (GCR);
- GMRES.

Let

$$V_k = [v_1, \cdots, v_k], \quad \tilde{H}_k = \begin{bmatrix} h_{1,1} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & \cdots & \cdots & h_{2,k} \\ : & : & : & : \\ \cdots & \cdots & h_{k,k-1} & h_{k,k} \\ 0 & \cdots & 0 & h_{k+1,k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times k}.$$

By Arnoldi algorithm, we have

$$AV_k = V_{k+1}\tilde{H}_k. \tag{4.14.8}$$

To solve the least square problem:

$$\min_{z \in K_k} \|r_o - Az\|_2 = \min_{z \in K_k} \|b - A(x_o + z)\|_2, \tag{4.14.9}$$

where $K_k =< r_o, Ar_o, \cdots, A^{k-1}r_o >=< v_1, \cdots, v_k >$ with $v_1 = \frac{r_o}{\|r_o\|_2}$. Set $z = V_k y$, the least square problem (4.14.9) is equivalent to

$$\min_{y\in\mathbb{R}^k} J(y) = \min_{y\in\mathbb{R}^k} \|\beta v_1 - AV_k y\|2, \quad \beta = \|r_o\|_2. \tag{4.14.10}$$

Using (4.14.8), we have

$$J(y) = \|V_{k+1}[\beta e_1 - \tilde{H}_k y]\|_2 = \|\beta e_1 - \tilde{H}_k y\|_2. \tag{4.14.11}$$

Hence, the solution of the least square (4.14.9) is

$$x_k = x_o + V_k y_k,$$

where y_k minimize the function $J(y)$ defined by (4.14.11) over $y \in \mathbb{R}^k$.

Algorithm 4.14.4 (GMRES algorithm)

Input: *choose x_0, compute $r_0 = b - Ax_0$ and $v_1 = r_0/\|r_0\|$;*
Output: *solution of linear system $Ax = b$.*
  Iterate $j = 1, 2, \cdots, k$,
    compute $h_{ij} = (Av_j, v_i)$ *for* $i = 1, 2, \cdots, j$,
      $\tilde{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i$,
      $h_{j+1,j} = \|\tilde{v}_{j+1}\|_2$,
      $v_{j+1} = \tilde{v}_{j+1}/h_{j+1,j}$.
  End;
  Form the solution:
      $x_k = x_0 + V_k y_k$, *where* y_k *minimizes* $J(y)$ *in* (4.14.11).

Difficulties: when k is increasing, storage for v_j, like k, the number of multiplications is like $\frac{1}{2}k^2N$.

Algorithm 4.14.5 (GMRES(m) algorithm)

Input: *choose x_0, compute $r_0 = b - Ax_0$ and $v_1 = r_0/\|r_0\|$;*
Output: *solution of linear system $Ax = b$.*
  Iterate $j = 1, 2, \cdots, m$,
    compute $h_{ij} = (Av_j, v_i)$ *for* $i = 1, 2, \cdots, j$,
      $\tilde{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i$,
      $h_{j+1,j} = \|\tilde{v}_{j+1}\|_2$,
      $v_{j+1} = \tilde{v}_{j+1}/h_{j+1,j}$.
  End;
  Form the solution:
      $x_m = x_0 + V_m y_m$, *where* y_m *minimizes* $\| \beta e_1 - \widetilde{H}_m y \|$ *for* $y \in \mathbb{R}^m$.
  Restart: *Compute $r_m = b - Ax_m$, if $\|r_m\|$ is small , then stop,*
      else , Compute $x_0 = x_m$ and $v_1 = r_m/ \| r_m \|$, GoTo Iterate *step.*

4.14.3 Practical Implementation: Consider QR factorization of $\widetilde{H}_k$

Consider the matrix $\widetilde{H}_k$, and let us suppose that we want to solve the least squares problem:

$$\min_{y\in\mathbb{R}^k} \parallel \beta e_1 - \widetilde{H}_k y \parallel_2$$

Assume Givens rotations F_i , $i = 1\ldots, j$ such that

$$F_j \cdots F_1 \widetilde{H}_j = F_j \ldots F_1 \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times \\ & \times & \times & \times \\ & & \times & \times \\ & & & \times \\ & & & 0 \end{bmatrix} \equiv R_j \in \mathbb{R}^{(j+1)\times j}.$$

In order to obtain R_{j+1} we must start by premultiptying the new column by the previous rotations.

$$\widetilde{H}_{j+1} = \left[\begin{array}{ccccc} \times & \times & \times & \times & + \\ \times & \times & \times & \times & + \\ 0 & \times & \times & \times & + \\ 0 & 0 & \times & \times & + \\ 0 & 0 & 0 & \times & + \\ \hline 0 & 0 & 0 & 0 & + \end{array}\right] \Rightarrow F_j \ldots F_i \widetilde{H}_{j+1} = \begin{bmatrix} \times & \times & \times & \times & + \\ & \times & \times & \times & + \\ & & \times & \times & + \\ & & & \times & + \\ & & & 0 & r \\ & & & 0 & h \end{bmatrix}$$

The principal upper $(j+1)\times j$ submatrix of the above matrix is nothing but R_j, and $h := h_{j+2,j+1}$ is not affected by the previous rotations. The next rotation F_{j+1} defined by

$$\begin{cases} c_{j+1} & \equiv & r/(r^2+h^2)^{1/2}, \\ s_{j+1} & = & -h/(r^2+h^2)^{1/2}. \end{cases}$$

Thus, after k steps of the above process, we have achieved

$$Q_k \widetilde{H}_k = R_k$$

where Q_k is a $(k+1)\times(k+1)$ unitary matrix and

$$J(y) = \parallel \beta e_1 - \widetilde{H}_k y \parallel = \parallel Q_k[\beta e_1 - \widetilde{H}_k y] \parallel = \parallel g_k - R_k y \parallel, \tag{4.14.12}$$

where $g_k \equiv Q_k \beta e_1$. Since the last row of R_k is a zero row, the minimization of (4.14.12) is achieved at $y_k = \widetilde{R}_k^{-1}\widetilde{g}_k$, where $\widetilde{R}_k$ and $\widetilde{g}_k$ are removed the last row of R_k and the last component of g_k, respectively.

Proposition 4.14.1 $\parallel r_k \parallel = \parallel b - Ax_k \parallel = \mid$ *The (k+1)-st component of* $g_k \mid$.

To avoid the extra computation needed to obtain x_k explicitly we suggest an efficient implementation of the last step of GMRES(m). To compute x_m we need to compute $\widetilde{H}_m$

and $v_1, \ldots, v_m$. Since $v_1, \cdots, v_m$ are known, we need to compute $h_{i,m}$, for $i = 1, \ldots, m+1$, of the form

$$\left[\begin{array}{cccc} h_{11} & \ldots & h_{1m-1} & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & h_{m,m-1} & h_{mm} \\ \hline & & 0 & h_{m+1,m} \end{array}\right]$$

with $h_{i,m} = (Av_m, v_i)$, for $i \leq m$. Here $h_{m+1,m}$ satisfies

$$h_{m+1,m}^2 = \| Av_m - \sum_{i=1}^{m} h_{im} v_i \|^2 = \| Av_m \|^2 - \sum_{i=1}^{m} h_{i,m}^2,$$

because

$$Av_m - \sum_{i=1}^{m} h_{im} v_i = h_{m+1,m} v_{m+1}, \qquad v_{m+1} \perp v_i, \quad \text{for} \quad i = 1, \ldots, m.$$

Now we will show how to compute $r_m = b - Ax_m$ from v_i's $i = 1, \ldots, m$ and Av_m. From (4.14.11) the residual vector can be expressed as

$$r_m = V_{m+1}[\beta e_1 - \widetilde{H}_m y_m].$$

Define $t \equiv [t_1, t_2, \ldots, t_{m+1}]^T \equiv \beta e_1 - \widetilde{H}_m y_m$. Then

$$\begin{aligned} v_m &= (\sum_{i=1}^{m} t_i v_i) + t_{m+1} v_{m+1} \\ &= (\sum_{i=1}^{m} t_i v_i) + t_{m+1} \frac{1}{h_{m+1,m}} [Av_m - \sum_{i=1}^{m} h_{i,m} v_i] \\ &= \frac{t_{m+1}}{h_{m+1,m}} Av_m + \sum_{i=1}^{m} (t_i - t_{m+1} h_{i,m}/h_{m+1,m}) v_i. \end{aligned}$$

Assume the first $m-1$ Arnoldi steps have been performed that the first $m-1$ columns of $\widetilde{H}_m$ and the first m vectors $v_i, i = 1, \ldots, m$ are available. Since we will not normalize v_i at every step, we do not have explicitly v_i but rather $w_i = \mu_i v_i$, μ_i are some known scaling coefficient (e.g., $\mu_i = \|v_i\|$). We have shown that r_m is a linear combination of Av_m and v_i's, $i = 1, \ldots, m$. Hence after m steps we do not need v_{m+1}. (Note that computing $\widetilde{v}_{m+1}$ and its norm costs $(2m+1)n$ multiplications. So elimination of its computation is a significant saving). So using $v_1, \ldots, v_m$ and Av_m we can compute restarting vector $v_1 := r_m / \| r_m \|$ and don't need to compute v_{m+1}. Then

$$r_m = \frac{t_{m+1}}{h_{m+1,m}} Av_m + \sum_{i=1}^{m} (t_i - t_{m+1} h_{i,m}/h_{m+1,m}) v_i.$$

By Proposition 4.14.1 it holds that $\| r_m \|_2 = |$ the $(k+1)$-st component of $g_k |$. So $v_1 := r_m / \| r_m \|_2$.

4.14.4 Theoretical Aspect of GMRES

GMRES CANNOT break down! GCR can break down when A is not positive real, i.e., $\frac{1}{2}(A+A^T)$ is not symmetric positive definite. We assume that the first m Arnoldi vectors can be constructed. That is, $h_{j+1,j} \neq 0$, for $j = 1, 2, \ldots, m$. In fact, if $h_{j+2,j+1} \neq 0$, the diagonal element $r_{j+1,j+1}$ of R_{j+1} satisfies

$$r_{j+1,j+1} = (c_{j+1}r - s_{j+1}h_{j+2,j+1}) = (r^2 + h_{j+2,j+1}^2)^{1/2} > 0.$$

Hence, the diagonal elements of R_m do not vanish and the least squares problem $J(y) = \min \|g_m - R_m y\|_2$ can be solved, establishing that the algorithm can not break down if $h_{j+1,j} \neq 0$, for $j = 1, \ldots, m$.

Thus the only possible potential difficulty is that during the Arnoldi process we encounter $h_{j+1,j} = 0$. From Arnoldi's algorithm it is easily seen that

(i) $AV_j = V_j H_j$ which means that K_j spanned by V_j is invariant. Note that if A is nonsingular then the eigenvalues of H_j are nonzero. $J(y)$ in (4.14.10) at the jth step becomes

$$J(y) = \|\beta v_1 - AV_j y\| = \|\beta v_1 - V_j H_j y\| = \|V_j[\beta e_1 - H_j y]\| = \|\beta e_1 - H_j y\|.$$

Since H_j is nonsingular, the above function is minimum for $y = H_j^{-1}\beta e_1$ and the corresponding minimum norm is zero, i.e., the solution x_j is exact.

Conversely, assume x_j is the exact solution and x_i, for $i = 1, \ldots, j-1$ are not, i.e. $r_j = 0$ but $r_i \neq 0$, for $i = 0, 1, \ldots, j-1$. From Proposition 4.14.1 we know that

$$\|r_j\| = s_j e_{j-1}^T g_{j-1} = s_j \|r_{j-1}\| = 0.$$

Then $s_j = 0$ ($\|r_{j-1}\| \neq 0$) which implies that $h_{j+1,j} = 0$, i.e., the algorithm breaks down and $\tilde{v}_{j+1} = 0$ which proves the result.

(ii) $\tilde{v}_{j+1} = 0$ and $\tilde{v}_i \neq 0$, for $i = 1, \ldots, j$ $\Leftrightarrow$ the degree of minimal polynomial of $r_0 = v_1$ is equal to j.

($\Leftarrow$) Assume that there exists a polynomial p_j of degree j such that $p_j(A)v_1 = 0$ and p_j is the polynomial of the lowest degree for which this is true. Therefore, $K_{j+1} =< v_1, Av_1, \cdots, A^j v_1 >= K_j$ so $\tilde{v}_{j+1} \in K_{j+1} = K_j$ and $\tilde{v}_{j+1} \perp K_j$, then $\tilde{v}_{j+1} = 0$. Moreover, if $\tilde{v}_i = 0$ for some $i \leq j$ then there is a polynomial p_i of degree i such that $p_i(A)v_i = 0$. This contradicts the minimality of p_j.

($\Rightarrow$) There is a polynomial p_j of degree j such that $p_j(A)v_1 = 0$ (by assumption $\tilde{v}_{j+1} = c, \tilde{v}_i \neq 0, i = 1, \ldots, j$). p_j is the polynomial of the lowest degree for which this is true. Otherwise, we have $\tilde{v}_i = 0$, for some $i < j+1$ by the first part of this proof. This is contradiction.

Proposition 4.14.2 *The solution x_j produced by GMRES at step j is exact which is equivalent to*

(i) The algorithm breaks down at step j,

(ii) $\tilde{v}_{j+1} = 0$,

(iii) $h_{j+1,j} = 0$,

(iv) The degree of the minimal polynomial of r_0 *is* j.

Corollary 4.14.1 *For an* $n \times n$ *problem GMRES terminates at most* n *steps.*

This uncommon type of breakdown is sometimes referred to as a "Lucky" breakdown is the context of the Lanczos algorithm.

Proposition 4.14.3 *Suppose that* A *is diagonalizable so that* $A = XDX^{-1}$ *and let*

$$\varepsilon^{(m)} = \min_{p \in P_m, p(0)=1} \max_{\lambda_i \in \sigma(A)} |p(\lambda_i)| .$$

Then

$$\|r_{m+1}\| \leq \kappa(X)\varepsilon^{(m)} \|r_0\| ,$$

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

When A is positive real with symmetric part M, it holds that

$$\|r_m\| \leq [1 - \alpha/\beta]^{m/2} \|r_0\| ,$$

where $\alpha = (\lambda_{\min}(M))^2$ and $\beta = \lambda_{\max}(A^T A)$.

This proves the convergence of GMRES(m) for all m, when A is positive real.

Theorem 4.14.2 *Assume* $\lambda_1, \ldots, \lambda_v$ *of* A *with positive(negative) real parts and the other eigenvalues enclosed in a circle centered at* C *with* $C > 0$ *and have radius* R *with* $C > R$. *Then*

$$\varepsilon^{(m)} \leq \left[\frac{R}{C}\right]^{m-\nu} \max_{j=\nu+1,\cdots,N} \prod_{i=1}^{\nu} \frac{|\lambda_i - \lambda_j|}{|\lambda_i|} \leq \left[\frac{D}{d}\right]^2 \left[\frac{R}{C}\right]^{m-\nu}$$

where

$$D = \max_{\substack{i=1,\cdots,\nu \\ j=\nu+1,\cdots,N}} |\lambda_i - \lambda_j| \quad \text{and} \quad d = \min_{i=1,\cdots,\nu} |\lambda_i| .$$

Proof: Consider $p(z) = r(z)q(z)$ where $r(z) = (1 - z/\lambda_1) \cdots (1 - z/\lambda_\nu)$ and $q(z)$ arbitrary polynomial of $deg \leq m - \nu$ such that $q(0) = 1$. Since $p(0) = 1$ and $p(\lambda_i) = 0$, for $i = 1, \ldots, \nu$, we have

$$\varepsilon^{(m)} \leq \max_{j=\nu+1,\cdots,N} |p(\lambda_j)| \leq \max_{j=\nu+1,\cdots,N} |r(\lambda_j)| \max_{j=\nu+1,\cdots,N} |q(\lambda_j)| .$$

It is easily seen that

$$\max_{j=\nu+1,\cdots,N} |r(\lambda_j)| = \max_{j=\nu+1,\cdots,N} \prod_{i=1}^{\nu} \frac{|\lambda_i - \lambda_j|}{|\lambda_i|} \leq \left[\frac{D}{d}\right]^{\nu} .$$

By maximin principle, the maximin of $|q(z)|$ for $z \in \{\lambda_j\}_{j=\nu+1}^{N}$ is no larger than its maximin over the circle that encloses that set. Taking $\sigma(z) = [(C - z)/C]^{m-\nu}$ whose maximin modulus on the circle is $(R/C)^{m-\nu}$ yields the desired result. ■

Corollary 4.14.2 *Under the assumptions of Proposition 4.14.3 and Theorem 4.14.2, GMRES(m) converges for any initial* x_0 *if*

$$m > \nu Log \left[\frac{DC}{dR}\kappa(X)^{1/\nu}\right] \Big/ Log \left[\frac{C}{R}\right] .$$

Part II

On the Numerical Solutions of Eigenvalue Problems

Chapter 5

The Unsymmetric Eigenvalue Problem

Generalized eigenvalue problem (GEVP):
Given $A, B \in \mathbb{C}^{n\times n}$. Determine $\lambda \in \mathbb{C}$ and $0 \neq x \in \mathbb{C}^n$ with $Ax = \lambda Bx$. λ is called an eigenvalue of the pencil $A - \lambda B$ (or *pair*(A, B)) and x is called an eigenvector corresponding to λ. λ is an eigenvalue of $A - \lambda B \iff \det(A - \lambda B) = 0$. ($\sigma(A, B) \equiv \{\lambda \in \mathbb{C} \mid \det(A - \lambda B) = 0\}$.)

Definition 5.0.2 *A pencil $A - \lambda B$ ($A, B \in \mathbb{R}^{m\times n}$) or a pair$(A, B)$ is called regular if that*

(i) *A and B are square matrices of order n, and*

(ii) $\det(A - \lambda B) \not\equiv 0$.

In all other case ($m \neq n$ or $m = n$ but $\det(A - \lambda B) \equiv 0$), the pencil is called singular.

Detailed algebraic structure of a pencil $A - \lambda B$ see Matrix theory II, chapter XII (Gantmacher 1959).

Eigenvalue Problem (EVP):
Special case in GEVP when $B = I$, we have $\lambda \in \mathbb{C}$ and $0 \neq x \in \mathbb{C}^n$ with $Ax = \lambda x$. λ is an eigenvalue of A and x is an eigenvector corresponding to λ.

Definition 5.0.3 (a) *$\sigma(A) = \{\lambda \in \mathbb{C} \mid \det(A - \lambda I) = 0\}$ is called the spectrum of A.*

(b) *$\rho(A) = \max\{|\ \lambda\ |: \lambda \in \sigma(A)\}$ is called the radius of $\sigma(A)$.*

(c) *$P(\lambda) = \det(\lambda I - A)$ is called the characteristic polynomial of A.*

$$\text{Let } P(\lambda) = \prod_{i=1}^{s} (\lambda - \lambda_i)^{m(\lambda_i)}, \quad \lambda_i \neq \lambda_j (i \neq j) \text{ and } \sum_{i=1}^{s} m(\lambda_i) = n.$$

Example 5.0.2 $A = \begin{bmatrix} 2 & 2 \\ 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Longrightarrow \det(A - \lambda B) = 2 - \lambda$ *and* $\sigma(A, B) = \{2\}$.

Example 5.0.3 $A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \Longrightarrow \det(A - \lambda B) = 3$ *and* $\sigma(A, B) = \emptyset$.

Example 5.0.4 $A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Longrightarrow \det(A - \lambda B) = 0$ *and* $\sigma(A, B) = \mathbb{C}$.

Example 5.0.5 $\det(\mu A - \lambda B) = (2\mu - \lambda)\mu$
$\mu = 1 : Ax = \lambda Bx \Longrightarrow \lambda = 2.$
$\lambda = 1 : Bx = \mu Ax \Longrightarrow \mu = 0, \mu = \frac{1}{2} \Longrightarrow \lambda = \infty, \lambda = 2.$
$\sigma(A, B) = \{2, \infty\}.$

Example 5.0.6 $\det(\mu A - \lambda B) = \mu \cdot 3\mu$
$\mu = 1 :$ *no solution for* λ.
$\lambda = 1 : Bx = \mu Ax \Longrightarrow \mu = 0, 0.$ *(multiple)*
$\sigma(A, B) = \{\infty, \infty\}.$

Let
$m(\lambda_i) :=$ algebraic multiplicity of λ_i.
$n(\lambda_i) := n - rank(A - \lambda_i I) =$ geometric multiplicity.
$1 \le n(\lambda_i) \le m(\lambda_i)$.
If for some i, $n(\lambda_i) < m(\lambda_i)$, then A is degenerated (defective). The following statements are equivalent:

(a) A is diagonalizable: There exists a nonsingular matrix T such that $T^{-1}AT = diag(\lambda_1, \cdots, \lambda_n)$.

(b) There are n linearly independent eigenvectors.

(c) A is nondefective, *i.e.* $\forall\ \lambda \in \sigma(A) \Longrightarrow m(\lambda) = n(\lambda)$.

If A is defective then eigenvector + principle vector $\Longrightarrow$ Jordan form.

Theorem 5.0.3 (Jordan decomposition) *If $A \in \mathbb{C}^{n\times n}$, then there exists a nonsingular $X \in \mathbb{C}^{n\times n}$, such that $X^{-1}AX = diag(J_1, \cdots, J_t)$,where*

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{bmatrix}$$

is $m_i \times m_i$ and $m_1 + \cdots + m_t = n$.

Theorem 5.0.4 (Schur decomposition) *If $A \in \mathbb{C}^{n\times n}$ then there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ such that $U^*AU (= U^{-1}AU)$ a upper triangular.*

- *A normal(i.e. $AA^* = A^*A$) $\Longleftrightarrow \exists$ unitary U such that $U^*AU = diag(\lambda_1, \cdots, \lambda_n)$, i.e. $Au_i = \lambda_i u_i, \quad u_i^* u_j = \delta_{ij}$.*

- *A hermitian(i.e. $A^* = A$) $\Longleftrightarrow$ A is normal and $\sigma(A) \subseteq \mathbb{R}$.*

- *A symmetric(i.e. $A^T = A, A \in \mathbb{R}^{n\times n}$) $\Longleftrightarrow \exists$ orthogonal U such that $U^TAU = diag(\lambda_1, \cdots, \lambda_n)$ and $\sigma(A) \subseteq \mathbb{R}$.*

5.1 Orthogonal Projections and C-S Decomposition

Definition 5.1.1 *Let $S \subseteq \mathbb{R}^n$ be a subspace, $P \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto S if*

$$\begin{cases} Range(P) = S, \\ P^2 = P, \\ P^T = P, \end{cases} \tag{5.1.1}$$

where $Range(P) = \mathcal{R}(P) = \{y \in \mathbb{R}^n \mid y = Px, \text{ for some } x \in \mathbb{R}^n\}$.

Remark 5.1.1 *If $x \in \mathbb{R}^n$, then $Px \in S$ and $(I - P)x \in S^\perp$.*

Example 5.1.1 *$P = vv^T/v^Tv$ is the orthogonal projection onto $S = span\{v\}, v \in \mathbb{R}^n$.*

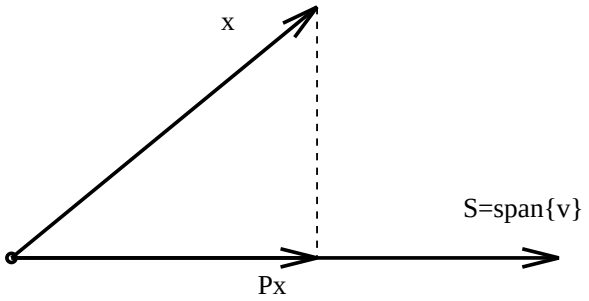


Figure 5.1: Orthogonal projection

Remark 5.1.2 **(i)** *If P_1 and P_2 are orthogonal projections, then for any $z \in \mathbb{R}^n$ we have*

$$\| (P_1 - P_2)z \|_2^2 = (P_1z)^T(I - P_2)z + (P_2z)^T(I - P_1)z. \tag{5.1.2}$$

If $\mathcal{R}(P_1) = \mathcal{R}(P_2) = S$ then the right-hand side of (5.1.2) is zero. Thus the orthogonal projection for a subspace is unique.

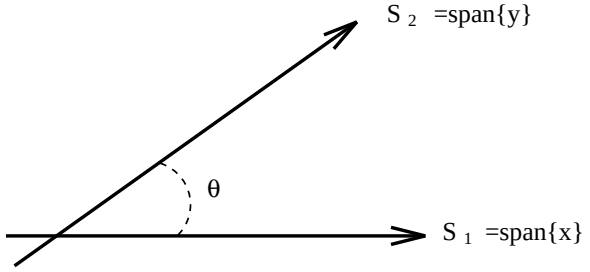
(ii) *If $V = [\mathbf{v}_1, \cdots, \mathbf{v}_k]$ is an orthogonal basis for S, then $P = VV^T$ is unique orthogonal projection onto S.*

Definition 5.1.2 *Suppose S_1 and S_2 are subspaces of $\mathbb{R}^n$ and $dim(S_1) = dim(S_2)$. We define the distance between S_1 and S_2 by*

$$dist(S_1, S_2) = \| P_1 - P_2 \|_2, \tag{5.1.3}$$

where P_i is the orthogonal projection onto S_i, $\quad i = 1, 2$.

Remark 5.1.3 *By considering the case of one-dimensional subspaces in $\mathbb{R}^2$, we obtain a geometrical interpretation of $dist(\cdot, \cdot)$. Suppose $S_1 = span\{x\}$ and $S_2 = span\{y\}$ and*



$\| x \|_2 = \| y \|_2 = 1$. *Assume that* $x^T y = \cos\theta, \theta \in [0, \frac{\pi}{2}]$. *It follows that the difference between the projections onto these spaces satisfies*

$$P_1 - P_2 = xx^T - yy^T = x[x - (y^T x)y]^T - [y - (x^T y)x]y^T.$$

If $\theta = 0 (\Rightarrow x = y)$, *then* $dist(S_1, S_2) = \| P_1 - P_2 \|_2 = \sin\theta = 0$.
If $\theta \neq 0$, *then*

$$U_x = [u_1, u_2] = [x, -[y - (y^T x)x]/\sin\theta]$$

and

$$V_x = [v_1, v_2] = [[x - (x^T y)y]/\sin\theta, y]$$

are defined and orthogonal. It follows that

$$P_1 - P_2 = U_x \; diag[\sin\theta, \sin\theta] \; V_x^T$$

is the SVD of $P_1 - P_2$. *Consequently,* $dist(S_1, S_2) = \sin\theta$, *the sine of the angle between the two subspaces.*

Theorem 5.1.1 (C-S Decomposition, Davis / Kahan(1970) or Stewart(1977)) *If* $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ *is orthogonal with* $Q_{11} \in \mathbb{R}^{k\times k}$ *and* $Q_{22} \in \mathbb{R}^{j\times j} (k \geq j)$, *then there exists orthogonal matrices* $U_1, V_1 \in \mathbb{R}^{k\times k}$ *and orthogonal matrices* $U_2, V_2 \in \mathbb{R}^{j\times j}$ *such that*

$$\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}^T \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} = \left[\begin{array}{cc|c} I & 0 & 0 \\ 0 & C & S \\ \hline 0 & -S & C \end{array}\right], \tag{5.1.4}$$

where

$$\begin{gathered} C = diag(c_1, \cdots, c_j), \quad c_i = \cos\theta_i, \\ S = diag(s_1, \cdots, s_j), \quad s_i = \sin\theta_i \\ and \; 0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_j \leq \tfrac{\pi}{2}. \end{gathered}$$

Lemma 5.1.1 *Let* $Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ *be orthogonal with* $Q_1 \in \mathbb{R}^{n\times n}$. *Then there are unitary matrices* U_1, U_2 *and* W *such that*

$$\begin{bmatrix} U_1^T & 0 \\ 0 & U_2^T \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} W = \begin{bmatrix} C \\ S \end{bmatrix}$$

where $C = diag(c_1, \cdots, c_j) \geq 0$, *and* $S = diag(s_1, \cdots, s_n) \geq 0$ *with* $c_i^2 + s_i^2 = 1, i = 1, \cdots, n$.

Proof: Let $U_1^T Q_1 W = C$ be the SVD of Q_1. Consider

$$\begin{bmatrix} U_1^T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} W = \begin{bmatrix} C \\ Q_2 W \end{bmatrix}$$

has orthogonal columns. Define $\tilde{Q}_2 \equiv Q_2 W$. Then $C^2 + \tilde{Q}_2^T \tilde{Q}_2 = I$ or $\tilde{Q}_2^T \tilde{Q}_2 = I - C^2$ diagonal, thus $\tilde{Q}_2^T \tilde{Q}_2$ is diagonal. Which means that the nonzero column of $\tilde{Q}_2$ are orthogonal to one another.If all the columns of $\tilde{Q}_2$ are nonzero, set $S^2 = \tilde{Q}_2^T \tilde{Q}_2$ and $U_2 = \tilde{Q}_2 S^{-1}$, then we have $U_2^T U_2 = I$ and $U_2^T \tilde{Q}_2 = S$. It follows the decomposition.

If $\tilde{Q}_2$ has zero columns, normalize the nonzero columns and replace the zero columns with an orthogonal basis for the orthogonal complement of the column space of $\tilde{Q}_2$. It is easily verified that U_2 so defined is orthogonal and $S \equiv U_2^T \tilde{Q}_2$ is diagonal. It also follows that decomposition. ■

Theorem 5.1.2 (C-S Decomposition) *Let the unitary matrix $W \in \mathbb{C}^{n\times n}$ be partitioned in the form $W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$, where $W_{11} \in \mathbb{C}^{r\times r}$ with $r \le \frac{n}{2}$. Then there exist unitary matrices $U = diag(\overbrace{U_1}^{r}, \overbrace{U_2}^{n-r})$ and $V = diag(\overbrace{V_1}^{r}, \overbrace{V_2}^{n-r})$ such that*

$$U^* W V = \begin{bmatrix} \Gamma & -\Sigma & 0 \\ \Sigma & \Gamma & 0 \\ 0 & 0 & I \end{bmatrix} \begin{matrix} \}r \\ \}r \\ \}n-2r \end{matrix}, \tag{5.1.5}$$

where $\Gamma = diag(\gamma_1, \cdots, \gamma_r) \ge 0$ and $\Sigma = diag(\sigma_1, \cdots, \sigma_r) \ge 0$ with $\gamma_i^2 + \sigma_i^2 = 1, i = 1, \cdots, r$.

Proof: Let $\Gamma = U_1^* W_{11} V_1$ be the SVD of W_{11} with the diagonal elements of Γ : $\gamma_1 \le \gamma_2 \le \cdots \le \gamma_k < 1 = \gamma_{k+1} = \cdots = \gamma_r$, *i.e.*

$$\Gamma = diag(\Gamma', I_{r-k}).$$

The matrix $\begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} V_1$ has orthogonal columns. Hence

$$I = \left[\begin{pmatrix} W_{11} \\ W_{21} \end{pmatrix} V_1 \right]^* \left[\begin{pmatrix} W_{11} \\ W_{21} \end{pmatrix} V_1 \right] = \Gamma^2 + (W_{21} V_1)^* (W_{21} V_1).$$

Since I and Γ^2 are diagonal, $(W_{21}V_1)^*(W_{21}V_1)$ is diagonal. So the columns of $W_{21}V_1$ are orthogonal. Since the ith diagonal of $I - \Gamma^2$ is the norm of the ith column of $W_{21}V_1$, only the first $k (k \le r \le n-r)$ columns of $W_{21}V_1$ are nonzero. Let $\hat{U}_2$ be unitary whose first k columns are the normalized columns of $W_{21}V_1$. Then

$$\hat{U}_2^* W_{21} V_1 = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix},$$

where $\Sigma = diag(\sigma_1, \cdots, \sigma_k, 0, \cdots, 0) \equiv \mathrm{diag}(\Sigma', 0)$, $\hat{U}_2 \in \mathbb{C}^{(n-r)\times(n-r)}$. Since

$$\mathrm{diag}(U_1, \hat{U}_2)^* \begin{pmatrix} W_{11} \\ W_{21} \end{pmatrix} V_1 = \begin{pmatrix} \Gamma \\ \Sigma \\ 0 \end{pmatrix}$$

has orthogonal (orthonormal) columns, we have $\gamma_i^2 + \sigma_i^2 = 1, i = 1, \cdots, r$. ($\Sigma'$ is nonsingular).

By the same argument as above : there is a unitary $V_2 \in \mathbb{C}^{(n-r)\times(n-r)}$ such that

$$U_1^* W_{12} V_2 = (T, 0),$$

where $T = diag(\tau_1, \cdots, \tau_r)$ and $\tau_i \leq 0$. Since $\gamma_i^2 + \tau_i^2 = 1$, it follows from $\gamma_i^2 + \sigma_i^2 = 1$ that $T = -\Sigma$. Set $\hat{U} = \mathrm{diag}(U_1, \hat{U}_2)$ and $V = diag(V_1, V_2)$. Then $X = \hat{U}^* W V$ can be partitioned in the form

$$X = \begin{bmatrix} \Gamma' & 0 & -\Sigma' & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \Sigma' & 0 & X_{33} & X_{34} & X_{35} \\ 0 & 0 & X_{43} & X_{44} & X_{45} \\ 0 & 0 & X_{53} & X_{54} & X_{55} \end{bmatrix} \begin{matrix} \}k \\ \}r-k \\ \}k \\ \}r-k \\ \}n-2r \end{matrix} .$$

Since columns 1 and 4 are orthogonal, it follows $\Sigma' X_{34} = 0$. Thus $X_{34} = 0$ (since Σ' nonsigular). Likewise $X_{35}, X_{43}, X_{53} = 0$. From the orthogonality of columns 1 and 3, it follows that $-\Gamma'\Sigma' + \Sigma' X_{33} = 0$, so $X_{33} = \Gamma'$. The matrix $\hat{U}_3 = \begin{bmatrix} X_{44} & X_{45} \\ X_{54} & X_{55} \end{bmatrix}$ is unitary. Set $U_2 = \mathrm{diag}(I_k, \hat{U}_3)\hat{U}_2$ and $U = diag(U_1, U_2)$. Then $U^H W V = diag(I_{r+k}, \hat{U}_3) X$ with

$$X = \begin{bmatrix} \Gamma' & 0 & -\Sigma' & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \Sigma' & 0 & \Gamma' & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}.$$

The theorem is proved. ■

Theorem 5.1.3 *Let $W = [W_1, W_2]$ and $Z = [Z_1, Z_2]$ be orthogonal, where W_1, $Z_1 \in \mathbb{R}^{n\times k}$ and $W_2, Z_2 \in \mathbb{R}^{n\times(n-k)}$. If $S_1 = R(W_1)$ and $S_2 = R(Z_1)$ then*

$$dist(S_1, S_2) = \sqrt{1 - \sigma_{min}^2(W_1^T Z_1)} \tag{5.1.6}$$

Proof: Let $Q = W^T Z$ and assume that $k \geq j = n - k$. Let the C-S decomposition of Q be given by (5.1.2), ($Q_{ij} = W_i^T Z_j, \quad i, j = 1, 2$). It follows that

$$\| W_1^T Z_2 \|_2 = \| W_2^T Z_1 \|_2 = s_j = \sqrt{1 - c_j^2} = \sqrt{1 - \sigma_{min}^2(W_1^T Z_1)}.$$

Since $W_1 W_1^T$ and $Z_1 Z_1^T$ are the orthogonal projections onto S_1 and S_2, respectively. We have

$$\begin{aligned} \mathrm{dist}(S_1, S_2) &= \| W_1 W_1^T - Z_1 Z_1^T \|_2 \\ &= \| W^T (W_1 W_1^T - Z_1 Z_1^T) Z \|_2 \\ &= \| \begin{bmatrix} 0 & W_1^T Z_2 \\ W_2^T Z_1 & 0 \end{bmatrix} \|_2 \\ &= s_j. \end{aligned}$$

If $k < j$, the above argument by setting $Q = [W_2, W_1]^T [Z_2, Z_1]$ and noting that

$$\sigma_{min}(W_2^T Z_1) = \sigma_{min}(W_1^T Z_2) = s_j.$$

■

5.2 Perturbation Theory

Theorem 5.2.1 (Gerschgorin Circle Theorem) *If $X^{-1}AX = D+F$, $D \equiv diag(d_1, \cdots, d_n)$ and F has zero diagonal entries, then $\sigma(A) \subset \bigcup_{i=1}^{n} D_i$, where*

$$D_i = \{z \in \mathbb{C} \mid |z - d_i| \leq \sum_{j=1, j\neq i}^{n} |f_{ij}|\}.$$

Proof: Suppose $\lambda \in \sigma(A)$ and assume without loss of generality that $\lambda \neq d_i$ for $i = 1, \cdots, n$. Since $(D - \lambda I) + F$ is singular, it follows that

$$1 \leq \| (D - \lambda I)^{-1} F \|_\infty = \sum_{j=1}^{n} |f_{kj}| / |d_k - \lambda|$$

for some $k (1 \leq k \leq n)$. But this implies that $\lambda \in D_k$. ■

Corollary 5.2.1 *If the union $M_1 = \bigcup_{j=1}^{k} D_{i_j}$ of k discs $D_{i_j}, j = 1, \cdots, k$, and the union M_2 of the remaining discs are disjoint, then M_1 contains exactly k eigenvalues of A and M_2 exactly $n - k$ eigenvalues.*

Proof: Let $B = X^{-1}AX = D + F$, for $t \in [0, 1]$. Let $B_t := D + tF$, then $B_0 = D, B_1 = B$. The eigenvalues of B_t are continuous functions of t. Applying Theorem 5.2.1 of Gerschgorin to B_t, one finds that for $t = 0$, there are exactly k eigenvalues of B_0 in M_1 and $n-k$ in M_2. (Counting multiple eigenvalues) Since for $0 \leq t \leq 1$ all eigenvalues of B_t likewise must lie in these discs, it follows for reasons of continuity that also k eigenvalues of A lie in M_1 and the remaining $n - k$ in M_2. ■

Remark 5.2.1 *Take $X = I, A = diag(A) + offdiag(A)$. Consider the transformation $A \longrightarrow \triangle^{-1}A\triangle$ with $\triangle = diag(\delta_1, \cdots, \delta_n)$. The Gerschgorin discs:*

$$D_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{\substack{k=1 \\ k\neq i}}^{n} \left| \frac{a_{ik}\delta_k}{\delta_i} \right| =: \rho_i\}.$$

Example 5.2.1 Let $A = \begin{bmatrix} 1 & \epsilon & \epsilon \\ \epsilon & 2 & \epsilon \\ \epsilon & \epsilon & 2 \end{bmatrix}$, $D_1 = \{z \mid |z - 1| \leq 2\epsilon\}$, $D_2 = D_3 = \{z \mid |z - 2| \leq 2\epsilon\}$, $0 < \epsilon \ll 1$. Transformation with $\triangle = diag(1, k\epsilon, k\epsilon), k > 0$ yields

$$\tilde{A} = \triangle^{-1}A\triangle = \begin{bmatrix} 1 & k\epsilon^2 & k\epsilon^2 \\ \frac{1}{k} & 2 & \epsilon \\ \frac{1}{k} & \epsilon & 2 \end{bmatrix}.$$

For $\tilde{A}$ we have $\rho_1 = 2k\epsilon^2, \rho_2 = \rho_3 = \frac{1}{k} + \epsilon$. The discs D_1 and $D_2 = D_3$ for $\tilde{A}$ are disjoint if

$$\rho_1 + \rho_2 = 2k\epsilon^2 + \tfrac{1}{k} + \epsilon < 1.$$

For this to be true we must clearly have $k > 1$. The optimal value $\tilde{k}$, for which D_1 and D_2(for $\tilde{A}$) touch one another, is obtained from $\rho_1 + \rho_2 = 1$. One finds

$$\tilde{k} = \frac{2}{1-\epsilon+\sqrt{(1-\epsilon)^2-8\epsilon^2}} = 1+\epsilon+O(\epsilon^2)$$

and thus $\rho_1 = 2\tilde{k}\epsilon^2 = 2\epsilon^2 + O(\epsilon^3)$. Through the transformation $A \longrightarrow \tilde{A}$ the radius ρ_1 of D_1 can thus be reduced from the initial 2ϵ to about $2\epsilon^2$.

Theorem 5.2.2 (Bauer-Fike) *If μ is an eigenvalue of $A+E \in \mathbb{C}^{n\times n}$ and $X^{-1}AX = D = diag(\lambda_1, \cdots, \lambda_n)$, then*

$$\min_{\lambda\in\sigma(A)} |\lambda-\mu| \leq \kappa_p(X) \parallel E \parallel_p,$$

where $\|\cdot\|_p$ is p-norm and $\kappa_p(X) = \parallel X \parallel_p \parallel X^{-1} \parallel_p$.

Proof: We need only consider the case $\mu \notin \sigma(A)$. If $X^{-1}(A+E-\mu I)X$ is singular, then so is $I + (D-\mu I)^{-1}(X^{-1}EX)$. Thus,

$$1 \leq \parallel (D-\mu I)^{-1}(X^{-1}EX) \parallel_p \leq \frac{1}{\min\limits_{\lambda\in\sigma(A)} |\lambda-\mu|} \parallel X \parallel_p \parallel E \parallel_p \parallel X^{-1} \parallel_p .$$

■

Theorem 5.2.3 *Let $Q^*AQ = D+N$ be a Schur decomposition of A with $D = diag(\lambda_1, \cdots, \lambda_n)$ and N strictly upper triangular, $N^n = 0$. If $\mu \in \sigma(A+E)$, then*

$$\min_{\lambda\in\sigma(A)} |\lambda-\mu| \leq \max\{\theta, \theta^{\frac{1}{n}}\},$$

where $\theta = \parallel E \parallel_2 \sum_{k=0}^{n-1} \parallel N \parallel_2^k$.

Proof: Define $\delta = \min_{\lambda\in\sigma(A)} |\lambda-\mu|$. The theorem is true if $\delta = 0$. If $\delta > 0$, then $I - (\mu I - A)^{-1}E$ is singular and we have

$$\begin{aligned} 1 &\leq \parallel (\mu I - A)^{-1}E \parallel_2 \\ &\leq \parallel (\mu I - A)^{-1} \parallel_2 \parallel E \parallel_2 \\ &= \parallel [(\mu I - D) - N]^{-1} \parallel_2 \parallel E \parallel_2 . \end{aligned}$$

Since $(\mu I - D)$ is diagonal it follows that $[(\mu I - D)^{-1}N]^n = 0$ and therefore

$$[(\mu I - D) - N]^{-1} = \sum_{k=0}^{n-1} [(\mu I - D)^{-1}N]^k (\mu I - D)^{-1}.$$

Hence we have

$$1 \leq \frac{\parallel E \parallel_2}{\delta} \max\{1, \frac{1}{\delta^{n-1}}\} \sum_{k=0}^{n-1} \parallel N \parallel_2^k,$$

from which the theorem readily follows. ■

Example 5.2.2 If $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 4.001 \end{bmatrix}$ and $E = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.001 & 0 & 0 \end{bmatrix}$. Then $\sigma(A+E) \cong \{1.0001, 4.0582, 3.9427\}$ and A's matrix of eigenvectors satisfies $\kappa_2(X) \cong 10^7$. The Bauer-Fike bound in Theorem 5.2.2 has order 10^4, but the Schur bound in Theorem 5.2.3 has order 10^0.

Theorems 5.2.2 and 5.2.3 each indicate potential eigenvalue sensitively if A is non-normal. Specifically, if $\kappa_2(X)$ and $\| N \|_2^{n-1}$ is large, then small changes in A can induce large change in the eigenvalues.

Example 5.2.3 If $A = \begin{bmatrix} 0 & I_9 \\ 0 & 0 \end{bmatrix}$ and $E = \begin{bmatrix} 0 & 0 \\ 10^{-10} & 0 \end{bmatrix}$, then for all $\lambda \in \sigma(A)$ and $\mu \in \sigma(A+E)$, $|\lambda - \mu| = 10^{\frac{-10}{10}}$. So a change of order 10^{-10} in A results in a change of order 10^{-1} in its eigenvalues.

Let λ be a simple eigenvalue of $A \in \mathbb{C}^{n\times n}$ and x and y satisfy $Ax = \lambda x$ and $y^*A = \lambda y^*$ with $\| x \|_2 = \| y \|_2 = 1$. Using classical results from **Function Theory**, it can be shown that there exists differentiable $x(\varepsilon)$ and $\lambda(\varepsilon)$ such that

$$(A + \varepsilon F)x(\varepsilon) = \lambda(\varepsilon)x(\varepsilon)$$

with $\| x(\varepsilon) \|_2 = 1$ and $\| F \|_2 \leq 1$, and such that $\lambda(0) = \lambda$ and $x(0) = x$. By differentiating and set $\varepsilon = 0$:

$$A\dot{x}(0) + Fx = \dot{\lambda}(0)x + \lambda\dot{x}(0).$$

Applying y^* to both sides and dividing by $y^*x \Longrightarrow$

$$f(x,y) = y^n + p_{n-1}(x)y^{n-1} + p_{n-2}(x)y^{n-2} + \cdots + p_1(x)y + p_0(x).$$

Fix x, then $f(x,y) = 0$ has n roots $y_1(x), \cdots, y_n(x)$. $f(0,y) = 0$ has n roots $y_1(0), \cdots, y_n(0)$.

Theorem 5.2.4 *Suppose $y_i(0)$ is a simple root of $f(0,y) = 0$, then there is $\delta_i > 0$ such that there is a simple root $y_i(x)$ of $f(x,y) = 0$ defined by*

$$y_i(x) = y_i(0) + p_{i1}x + p_{i2}x^2 + \cdots, \quad \textit{(may terminate!)}$$

where the series is convergent for $|x| < \delta_i$. ($y_i(x) \longrightarrow y_i(0)$ as $x \longrightarrow 0$).

Theorem 5.2.5 *If $y_1(0) = \cdots = y_m(0)$ is a root of multiplicity m of $f(0,y) = 0$, then there exists $\delta > 0$ such that there are exactly m zeros of $f(x,y) = 0$ when $|x| < \delta$ having the following properties:*

(a) $\sum_{i=1}^{r} m_i = m, \quad m_i \geq 0$. *The m roots fall into r groups.*

(b) *Those roots in the group of m_i are m_i values of a series*

$$y_1(0) + p_{i1}z + p_{i2}z^2 + \cdots$$

corresponding to the m_i different values of z defined by $z = x^{\frac{1}{m_i}}$.

Let λ_1 be a simple root of A and x_1 be the corresponding eigenvector. Since λ_1 is simple, $(A-\lambda_1 I)$ has at least one nonzero minor of order $n-1$. Suppose this lies in the first $(n-1)$ rows of $(A-\lambda_1 I)$. Take $x_1 = (A_{n1}, A_{n2}, \cdots, A_{nn})$. Then

$$(A-\lambda_1 I)\begin{pmatrix} A_{n1} \\ A_{n2} \\ \vdots \\ A_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

since $\sum_{j=1}^{n} a_{nj}A_{nj} = \det(A-\lambda_1 I) = 0$. Here A_{ni} is the cofactor of a_{ni}, hence it is a polynomial in λ_1 of degree not greater than $(n-1)$.

Let $\lambda_1(\varepsilon)$ be the simple eigenvalue of $A+\varepsilon F$ and $x_1(\varepsilon)$ be the corresponding eigenvector. Then the elements of $x_1(\varepsilon)$ are the polynomial in $\lambda_1(\varepsilon)$ and ε. Since the power series for $\lambda_1(\varepsilon)$ is convergent for small ε, so $x_1(\varepsilon) = x_1 + \varepsilon z_1 + \varepsilon^2 z_2 + \cdots$ is a convergent power series $\left|\dot{\lambda}(0)\right| = \dfrac{|y^*Fx|}{|y^*x|} \leq \dfrac{1}{|y^*x|}$. The upper bound is attained if $F = yx^*$. We refer to the reciprocal of $s(\lambda) \equiv |y^*x|$ as the condition number of the eigenvalue λ.

$\lambda(\varepsilon) = \lambda(0) + \dot{\lambda}(0)\varepsilon + O(\varepsilon^2)$, an eigenvalue λ may be perturbed by an amount $\dfrac{\varepsilon}{s(\lambda)}$, if $s(\lambda)$ is small then λ is appropriately regarded as ill-conditioned. Note that $s(\lambda)$ is the cosine of the angle between the left and right eigenvectors associated with λ and is *unique only* if λ is simple. A small $s(\lambda)$ implies that A is near a matrix having a multiple eigenvalue. In particular, if λ is distinct and $s(\lambda) < 1$, then there exists an E such that λ is a repeated eigenvalue of $A+E$ and

$$\| E \|_2 \leq \frac{s(\lambda)}{\sqrt{1-s^2(\lambda)}},$$

this is proved in Wilkinson(1972).

Example 5.2.4 If $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 4.001 \end{bmatrix}$ and $E = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.001 & 0 & 0 \end{bmatrix}$. Then $\sigma(A+E) \cong \{1.0001, 4.0582, 3.9427\}$ and $s(1) \cong 0.79 \times 10^0, s(4) = 0.16 \times 10^{-3}, s(4.001) \cong 0.16 \times 10^{-3}$. Observe that $\| E \|_2 / s(\lambda)$ is a good estimate of the perturbation that each eigenvalue undergoes.

If λ is a repeated eigenvalue, then the eigenvalue sensitivity question is more complicated. For example $A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$ and $F = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ then $\sigma(A+\varepsilon F) = \{1 \pm \sqrt{\varepsilon a}\}$. Note that if $a \neq 0$ then the eigenvalues of $A+\varepsilon F$ are not differentiable at zero, their rate of change at the origin is infinite. In general, if λ is a detective eigenvalue of A, then $O(\varepsilon)$ perturbations in A result in $O(\varepsilon^{\frac{1}{p}})$ perturbations in λ where $p \geq 2$ (see Wilkinson AEP pp.77 for a more detailed discussion).

We now consider the perturbations of invariant subspaces. Assume $A \in \mathbb{C}^{n\times n}$ has distinct eigenvalues $\lambda_1, \cdots, \lambda_n$ and $\| F \|_2 = 1$. We have

$$(A+\varepsilon F)x_k(\varepsilon) = \lambda_k(\varepsilon)x_k(\varepsilon), \quad \| x_k(\varepsilon) \|_2 = 1,$$

and

$$y_k^*(\varepsilon)(A+\varepsilon F) = \lambda_k(\varepsilon)y_k^*(\varepsilon), \quad \| y_k(\varepsilon) \|_2 = 1,$$

for $k = 1, \cdots, n$, where each $\lambda_k(\varepsilon)$,$x_k(\varepsilon)$ and $y_k(\varepsilon)$ are differentiable. Set $\varepsilon = 0$:

$$A\dot{x}_k(0) + Fx_k = \dot{\lambda}_k(0)x_k + \lambda_k\dot{x}_k(0),$$

where $\lambda_k = \lambda_k(0)$ and $x_k = x_k(0)$. Since $\{x_i\}_{i=1}^n$ linearly independent, write $\dot{x}_k(0) = \sum_{i=1}^n a_i x_i$, so we have

$$\sum_{\substack{i=1 \\ i\neq k}}^{n} a_i(\lambda_i - \lambda_k)x_i + Fx_k = \dot{\lambda}_k(0)x_k.$$

But $y_i^*(0)x_k = y_i^*x_k = 0$, for $i \neq k$ and thus

$$a_i = y_i^*Fx_k/[(\lambda_k - \lambda_i)y_i^*x_i], \quad i \neq k.$$

Hence the Taylor expansion for $x_k(\varepsilon)$ is

$$x_k(\varepsilon) = x_k + \varepsilon\sum_{\substack{i=1 \\ i\neq k}}^{n} \left\{ \frac{y_i^*Fx_k}{(\lambda_k - \lambda_i)y_i^*x_i} \right\} x_i + O(\varepsilon^2).$$

Thus the sensitivity of x_k depends upon eigenvalue sensitivity and the separation of λ_k from the other eigenvalues.

Example 5.2.5 If $A = \begin{bmatrix} 1.01 & 0.01 \\ 0.00 & 0.99 \end{bmatrix}$, then $\lambda = 0.99$ has Condition $\dfrac{1}{s(0.99)} \cong 1.118$ and associated eigenvector $x = (0.4472, -8.944)^T$. On the other hand, $\tilde{\lambda} = 1.00$ of the "nearby" matrix $A + E = \begin{bmatrix} 1.01 & 0.01 \\ 0.00 & 1.00 \end{bmatrix}$ has an eigenvector $\tilde{x} = (0.7071, -0.7071)^T$.

Suppose

$$Q^*AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{matrix} \}p \\ \}q = n-p \end{matrix} \tag{5.2.1}$$

is a Schur decomposition of A with

$$Q = [\underbrace{Q_1}_{p}, \underbrace{Q_2}_{n-p}]. \tag{5.2.2}$$

Definition 5.2.1 *We define the separation between T_{11} and T_{22} by*

$$sep_F(T_{11}, T_{22}) = \min_{Z\neq 0} \frac{\| T_{11}Z - ZT_{22} \|_F}{\| Z \|_F}.$$

Definition 5.2.2 *Let X be a subspace of $\mathbb{C}^n$, X is called an invariant subspace of $A \in \mathbb{C}^{n\times n}$, if $AX \subset X$ (i.e. $x \in X \Longrightarrow Ax \in X$).*

Theorem 5.2.6 *$A \in \mathbb{C}^{n\times n}, V \in \mathbb{C}^{n\times r}$ and $rank(V) = r$, then there are equivalent:*

(a) *there exists* $S \in \mathbb{C}^{r\times r}$ *such that* $AV = VS$.

(b) $R(V)$ *is an invariant subspace of* A.

Proof: Trivial! ■

Remark 5.2.2 **(a)** *If* $Sz = \mu z, z \neq 0$ *then* μ *is eigenvalue of* A *with eigenvector* Vz.

(b) *If* V *is a basis of* X, *then* $\tilde{V} = V(V^*V)^{-\frac{1}{2}}$ *is an orthogonal basis of* X.

Theorem 5.2.7 $A \in \mathbb{C}^{n\times n}, Q = (Q_1, Q_2)$ *orthogonal, then there are equivalent:*

(a) *If* $Q^*AQ = B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$, *then* $B_{21} = 0$.

(b) $R(Q_1)$ *is an invariant subspace of* A.

Proof: $Q^*AQ = B \iff AQ = QB = (Q_1, Q_2) \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$. Thus $AQ_1 = Q_1B_{11} + Q_2B_{21}$.

(a) $B_{21} = 0$, then $AQ_1 = Q_1B_{11}$, so $R(Q_1)$ is an invariant subspace of A (from Theorem 5.2.6).

(b) $R(Q_1)$ is invariant subspace. There exists S such that $AQ_1 = Q_1S = Q_1B_{11}+Q_2B_{21}$. Multiply with Q_1^*, then

$$S = Q_1^*Q_1S = Q_1^*Q_1B_{11} + Q_1^*Q_2B_{21}.$$

So $S = B_{11} \Longrightarrow Q_2B_{21} = 0 \Longrightarrow Q_2^*Q_2B_{21} = 0 \Longrightarrow B_{21} = 0$. ■

Theorem 5.2.8 *Suppose (5.2.1) and (5.2.2) hold and for* $E \in \mathbb{C}^{n\times n}$ *we partition* Q^*EQ *as follows:*

$$Q^*EQ = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

with $E_{11} \in \mathbb{R}^{p\times p}$ *and* $E_{22} \in \mathbb{R}^{(n-p)\times(n-p)}$. *If*

$$\delta = sep_2(T_{11}, T_{22}) - \| E_{11} \|_2 - \| E_{22} \|_2 > 0$$

and

$$\| E_{21} \|_2 (\| T_{12} \|_2 + \| E_{12} \|_2) \leq \delta^2/4.$$

Then there exists $P \in \mathbb{C}^{(n-k)\times k}$ *such that*

$$\| P \|_2 \leq 2 \| E_{21} \|_2 / \delta$$

and such that the column of $\tilde{Q}_1 = (Q_1 + Q_2P)(I + P^*P)^{-\frac{1}{2}}$ *form an orthonormal basis for a invariant subspace of* $A + E$.*(See Stewart 1973).*

Lemma 5.2.1 *Let* $\{s_m\}$ *and* $\{p_m\}$ *be two sequence defined by*

$$s_{m+1} = s_m/(1 - 2\eta p_m s_m), \quad p_{m+1} = \eta p_m^2 s_{m+1}, \quad m = 0, 1, 2, \cdots \tag{5.2.3}$$

and

$$s_0 = \sigma, \quad p_0 = \sigma\gamma \tag{5.2.4}$$

satisfying

$$4\eta\sigma^2\gamma < 1. \quad (Here \quad \sigma, \eta, \gamma > 0) \tag{5.2.5}$$

Then $\{s_m\}$ *is monotonic increasing and bounded above;* $\{p_m\}$ *is monotonic decreasing, converges quadratically to zero.*

Proof: Let

$$x_m = s_m p_m, \quad m = 0, 1, 2, \cdots . \tag{5.2.6}$$

From (5.2.3) we have

$$x_{m+1} = s_{m+1}p_{m+1} = \eta p_m^2 s_m^2/(1 - 2\eta p_m s_m)^2 = \eta x_m^2/(1 - 2\eta x_m)^2, \tag{5.2.7}$$

(5.2.5) can be written as

$$0 < x_0 < \frac{1}{4\eta}. \quad (since \quad x_0 = s_0 p_0 = \sigma^2\gamma < \frac{1}{4\eta}) \tag{5.2.8}$$

Consider

$$x = f(x), \quad f(x) = \eta x^2/(1 - 2\eta x)^2, \quad x \geq 0. \tag{5.2.9}$$

By

$$\frac{df(x)}{dx} = \frac{2\eta x}{(1 - 2\eta x)^3},$$

we know that $f(x)$ is differentiable and monotonic increasing in $[0, 1/2\eta)$, and $\frac{df(x)}{dx} \mid_{x=0} = 0$: The equation (5.2.9) has zeros 0 and $1/4\eta$ in $[0, 1/2\eta)$. Under Condition (5.2.8) the iteration x_m as in (5.2.7) must be monotone decreasing converges quadratically to zero. (*Issacson &Keller "Analysis of Num. Method 1996, Chapter 3 §1.*) Thus

$$\frac{s_{m+1}}{s_m} = \frac{1}{1 - 2\eta x_m} = 1 + \frac{2\eta x_m}{1 - 2\eta x_m} = 1 + t_m,$$

where t_m is monotone decreasing, converges quadratically to zero, hence

$$s_{m+1} = s_0 \prod_{j=0}^{m} \frac{s_{j+1}}{s_j} = s_0 \prod_{j=0}^{m} (1 + t_j)$$

monotone increasing, and converges to $s_0 \prod_{j=0}^{\infty} (1 + t_j) < \infty$, so $p_m = \frac{x_m}{s_m}$ monotone decreasing, and quadratically convergent to zero. ■

Theorem 5.2.9 *Let*

$$PA_{12}P + PA_{11} - A_{22}P - A_{21} = 0 \tag{5.2.10}$$

be the quadratic matrix equation in $P \in \mathbb{C}^{(n-l)\times l} \quad (1 \le l \le n)$, *where*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A, \quad \sigma(A_{11}) \bigcap \sigma(A_{22}) = \emptyset.$$

Define operator T *by:*

$$TQ \equiv QA_{11} - A_{22}Q, \quad Q \in \mathbb{C}^{(n-l)\times l}. \tag{5.2.11}$$

Let

$$\eta = \| A_{12} \|, \quad \gamma = \| A_{21} \| \tag{5.2.12}$$

and

$$\sigma = \| T^{-1} \| = \sup_{\|P\|=1} \| T^{-1}P \| . \tag{5.2.13}$$

If

$$4\eta\sigma^2\gamma < 1, \tag{5.2.14}$$

then according to the following iteration, we can get a solution P *of (5.2.10) satisfying*

$$\| P \| \le 2\sigma\gamma, \tag{5.2.15}$$

and this iteration is quadratic convergence.

Iteration: *Let* $A_m = \begin{bmatrix} A_{11}^{(m)} & A_{12}^{(m)} \\ A_{21}^{(m)} & A_{22}^{(m)} \end{bmatrix}, \quad A_0 = A.$

(i) *Solve*

$$T_m P_m \equiv P_m A_{11}^{(m)} - A_{22}^{(m)} P_m = A_{21}^{(m)} \tag{5.2.16}$$

and get $P_m \in \mathbb{C}^{(n-l)\times l}$;

(ii) *Compute*

$$\begin{aligned} A_{11}^{(m+1)} &= A_{11}^{(m)} + A_{12}P_m, \\ A_{22}^{(m+1)} &= A_{22}^{(m)} - P_m A_{12}, \\ A_{21}^{(m+1)} &= -P_m A_{12} P_m. \end{aligned}$$

Goto (i), *solve* P_{m+1}.

Then

$$P = \lim_{m\to\infty} \sum_{i=0}^{m} P_i \tag{5.2.17}$$

is a solution of (5.2.10) and satisfies (5.2.15).

Proof: (a) Prove that for $m = 0, 1, 2, \cdots$, T_m^{-1} exist: denote

$$\| T_m^{-1} \| = \sigma_m, \quad (T = T_0, \quad \sigma = \sigma_0), \tag{5.2.18}$$

then

$$4 \| A_{12} \| \| P_m \| \sigma_m < 1. \tag{5.2.19}$$

By induction, $m = 0$, from $\sigma(A_{11}) \bigcap \sigma(A_{22}) = \emptyset$ we have $T_0 = T$ is nonsingular. From (5.2.12)-(5.2.14) it holds

$$4 \| A_{12} \| \| P_0 \| \sigma_0 = 4\eta \| T^{-1} A_{21} \| \sigma \leq 4\eta\sigma^2\gamma < 1.$$

Suppose T_m^{-1} exists, and (5.2.19) holds, prove that T_{m+1}^{-1} exists and

$$4 \| A_{12} \| \| P_{m+1} \| \sigma_{m+1} < 1.$$

From the definition

$$\text{sep}(A_{11}, A_{22}) = \inf_{\|Q\|=1} \| QA_{11} - A_{22}Q \|$$

and the existence of T^{-1} follows $\text{sep}(A_{11}, A_{22}) = \| T^{-1} \|^{-1} = \sigma^{-1}$, and by the perturbation property of "sep" follows

$$\begin{aligned} &\text{sep}(A_{11}^{(m+1)}, A_{22}^{(m+1)}) = \text{sep}(A_{11}^{(m)} + A_{12}P_m, A_{22}^{(m)} - P_m A_{12}) \\ &\geq \text{sep}(A_{11}^{(m)}, A_{22}^{(m)}) - \| A_{12}P_m \| - \| P_m A_{12} \| \\ &\geq \frac{1 - 2 \| A_{12} \| \| P_m \| \sigma_m}{\sigma_m} > 0. \end{aligned} \tag{5.2.20}$$

From

$$\text{sep}(A_{11}, A_{22}) \leq \min\{|\lambda_1 - \lambda_2| : \lambda_1 \in \sigma(A_{11}), \lambda_2 \in \sigma(A_{22})\}.$$

We have $\sigma(A_{11}^{(m+1)}) \bigcap \sigma(A_{22}^{(m+1)}) = \emptyset$, hence T_{m+1}^{-1} exists and

$$\text{sep}(A_{11}^{(m+1)}, A_{22}^{(m+1)}) = \| T_{m+1}^{-1} \|^{-1} = \sigma_{m+1}^{-1}.$$

From (5.2.20) it follows

$$\sigma_{m+1} \leq \frac{\sigma_m}{1 - 2 \| A_{12} \| \| P_m \| \sigma_m}. \tag{5.2.21}$$

Substitute (5.2.19) into (5.2.21), we get $\sigma_{m+1} \leq 2\sigma_m$, and

$$\| P_{m+1} \| \leq \| T_{m+1}^{-1} \| \| A_{21}^{m+1} \| \leq \sigma_{m+1} \| P_m \|^2 \| A_{12} \| < \frac{1}{2} \| P_m \| .$$

Hence

$$2 \| A_{12} \| \| P_{m+1} \| \sigma_{m+1} \leq 2 \| A_{12} \| \| P_m \| \sigma_m < 1/2.$$

This proved that T_m^{-1} exists for all $m = 0, 1, 2, \cdots$ and (5.2.19) holds.

(b) Prove $\| P_m \|$ is quadratic convergence to zero. Construct sequences $\{q_m\}, \{s_m\}, \{p_m\}$ satisfying

$$\| A_{21}^{(m)} \| \leq q_m, \quad \sigma_m \leq s_m, \quad \| P_m \| \leq p_m. \tag{5.2.22}$$

From

$$A_{21}^{(m+1)} = -P_m A_{12} P_m \tag{5.2.23}$$

follows

$$\| A_{21}^{(m+1)} \| \leq \| A_{12} \| \| P_m \|^2 \leq \eta p_m^2. \tag{5.2.24}$$

Define $\{q_m\}$ by

$$q_{m+1} = \eta p_m^2, \quad q_0 = \gamma; \tag{5.2.25}$$

From (5.2.21) we have

$$\sigma_{m+1} \leq \frac{s_m}{1 - 2\eta p_m s_m}. \tag{5.2.26}$$

Define $\{s_m\}$ by

$$s_{m+1} = \frac{s_m}{1 - 2\eta p_m s_m}, \quad s_0 = \sigma; \tag{5.2.27}$$

From (5.2.16) we have

$$\| P_m \| \leq \| T_m^{-1} \| \| A_{21}^{(m)} \| = \sigma_m \| A_{21}^{(m)} \| \leq s_m q_m.$$

Define $\{p_m\}$ by

$$p_{m+1} = s_{m+1} q_{m+1} = \eta p_m^2 s_{m+1}, \quad p_0 = \sigma\gamma. \tag{5.2.28}$$

By Lemma 5.2.1 follows that $\{p_m\} \searrow 0$ monotone and form (5.2.22) follows that $\| P_m \| \longrightarrow 0$ quadratically.

(c) Prove $P^{(m)} \longrightarrow P$ and (5.2.15) holds. According to the method as in Lemma 5.2.1. Construct $\{x_m\}$ (see (5.2.6),(5.2.7)), that is

$$x_{m+1} = \frac{\eta x_m^2}{(1 - 2\eta x_m)^2}, \quad s_{m+1} = \frac{s_m}{1 - 2\eta x_m} \tag{5.2.29}$$

and then

$$p_{m+1} = \frac{x_{m+1}}{s_{m+1}} = \frac{\eta x_m}{1 - 2\eta x_m} p_m. \tag{5.2.30}$$

By induction! For all $m = 1, 2, \cdots$ we have

$$p_m < \frac{1}{2} p_{m-1}, \quad x_m < \frac{1}{4\eta}. \tag{5.2.31}$$

In fact, substitute

$$\frac{\eta x_0}{1 - 2\eta x_0} = \frac{\eta \sigma^2 \gamma}{1 - 2\eta\sigma^2\gamma} < \frac{1}{2} \tag{5.2.32}$$

into (5.2.30) and get $p_1 < \frac{1}{2} p_0$; From (5.2.29) and (5.2.32) it follows that

$$x_1 = \frac{1}{\eta} \left(\frac{\eta x_0}{1 - 2\eta x_0} \right)^2 < \frac{1}{4\eta}.$$

For $m = 1$, (5.2.31) holds. Suppose for m (5.2.31) holds, form (5.2.30), we have

$$p_{m+1} < \frac{1}{2} p_m;$$

by (5.2.29) it holds $x_{m+1} = \frac{1}{\eta}\left(\frac{\eta x_m}{1-2\eta x_m}\right)^2 < \frac{1}{4\eta}$, that is (5.2.31) holds for $m+1$. Hence (5.2.31) holds for all nature number m. Therefore $p_m < p_0/2^m, \quad m = 1, 2, \cdots$, hence $p^{(m)}$ converges, where

$$p^{(m)} = \sum_{i=0}^{m} p_i < \left(\sum_{i=0}^{m} \frac{1}{2^i}\right) p_0 = 2\left(1 - \frac{1}{2^{m+1}}\right) p_0. \tag{5.2.33}$$

Let

$$P^{(m)} = \sum_{i=0}^{m} P_i.$$

From (5.2.22),(5.2.28) and (5.2.33) follows that

$$\| P^{(m)} \| \leq \sum_{i=0}^{m} \| P_i \| \leq \sum_{i=0}^{m} p_i < 2\left(1 - \frac{1}{2^{m+1}}\right) p_0 = 2\left(1 - \frac{1}{2_{m+1}}\right) \sigma\gamma.$$

Let $m \longrightarrow \infty$, then (5.2.15) holds. By (b) the limit matrix P as in (5.2.17) is quadratic convergence. ■

Theorem 5.2.10 *Let $A, E \in \mathbb{C}^{n\times n}, \quad Z_1 \in \mathbb{C}^{n\times l}$ be the eigenmatrix of A corresponding to $A_{11} \in \mathbb{C}^{l\times l}$ (i.e. $AZ_1 = Z_1 A_{11}$) and $Z_1^H Z_1 = I, \quad 1 \leq l \leq n$. Let $Z = (Z_1, Z_2)$ be unitary. Denote*

$$Z^* A Z = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Z^* E Z = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}.$$

Define T as in (5.2.11). Suppose $\sigma(A_{11}) \bigcap \sigma(A_{22}) = \emptyset$ and $\| T^{-1} \| (\| E_{11} \| + \| E_{22} \|) < 1$. Let

$$\tilde{\sigma} = \frac{\| T^{-1} \|}{1 - \| T^{-1} \| (\| E_{11} \| + \| E_{22} \|)}, \quad \tilde{\eta} = \| A_{12} \| + \| E_{12} \|, \quad \tilde{\gamma} = \| E_{21} \|. \tag{5.2.34}$$

If

$$4\tilde{\eta}\tilde{\sigma}^2\tilde{\gamma} < 1, \tag{5.2.35}$$

then there exists $P \in \mathbb{C}^{(n-l)\times l}$ with $\| P \| \leq 2\tilde{\sigma}\tilde{\gamma}$ such that

$$\tilde{Z}_1 = Z_1 + Z_2 P \in \mathbb{C}^{n\times l} \tag{5.2.36}$$

is the eigenmatrix of $\tilde{A} = A + E$ corresponding to $\tilde{A}' = A_{11} + E_{11} + (A_{12} + E_{12})P$.

Proof: Prove that there exists $L = \begin{bmatrix} I_l & 0 \\ P & I_{n-l} \end{bmatrix}$ with $\| P \| \leq 2\tilde{\sigma}\tilde{\gamma}$ such that

$$L^{-1} \begin{bmatrix} A_{11}+E_{11} & A_{12}+E_{12} \\ E_{21} & A_{22}+E_{22} \end{bmatrix} L = \begin{bmatrix} \tilde{A}'_{11} & * \\ 0 & * \end{bmatrix}. \tag{5.2.37}$$

This is resulted from solving the following equation

$$P(A_{12}+E_{12})P + P(A_{11}+E_{11}) - (A_{22}+E_{22})P - E_{21} = 0. \tag{5.2.38}$$

Let

$$\tilde{T}P \equiv P(A_{11}+E_{11}) - (A_{22}+E_{22})P.$$

By (5.2.34),(5.2.35) and

$$\begin{aligned} \| \tilde{T}^{-1} \| &= \{ \inf_{\|P\|=1} \| P(A_{11}+E_{11}) - (A_{22}+E_{22})P \| \}^{-1} \\ &\leq \{ \inf_{\|P\|=1} \| PA_{11} - A_{22}P \| - \sup_{\|P\|=1} \| PE_{11} - E_{22}P \| \}^{-1} \\ &\leq \frac{\| T^{-1} \|}{1 - \| T^{-1} \| (\| E_{11} \| + \| E_{22} \|)} = \tilde{\sigma}, \end{aligned}$$

we have

$$4 \| (A_{12}+E_{12}) \| \| \tilde{T}^{-1} \|^2 \| E_{21} \| \leq 4\tilde{\eta}\tilde{\sigma}^2\tilde{\gamma} < 1.$$

Because the condition (5.2.14) in Theorem 5.2.9 is satisfied, by Theorem 5.2.9, the equation (5.2.38) has a solution P satisfying $\| P \| \leq 2\tilde{\sigma}\tilde{\gamma}$. Then it follows the result from (5.2.37). ■

Remark 5.2.3 *Normalized* $Z_1 + Z_2P \longrightarrow (Z_1 + Z_2P)(I + P^HP)^{\frac{-1}{2}}$. *Consider*

$$\begin{aligned} & dist(Z_1, (Z_1 + Z_2P)(I + P^HP)^{\frac{-1}{2}}) \\ &= \sqrt{1 - \sigma^2_{\min}[Z_1^H(Z_1 + Z_2P)(I + P^HP)^{\frac{-1}{2}}]} \\ &= \sqrt{1 - \sigma^2_{\min}[(I + P^HP)^{\frac{-1}{2}}]} \\ &= \sqrt{1 - [\sigma_{\max}(I + P^HP)]^{-1}} \\ &\leq \sqrt{1 - \frac{1}{1 + \| P \|_2^2}} \\ &= \frac{\| P \|_2}{\sqrt{1 + \| P \|_2^2}}. \end{aligned}$$

Example 5.2.6 Let $n = 3, l = 2, k = 1$,

$$A = \left[\begin{array}{cc|c} 6 & -1 & 1 \\ 1 & 4 & 0 \\ \hline 0 & 0 & 1 \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad E = \left[\begin{array}{cc|c} 0.5 & -0.1 & 0.3 \\ -0.4 & 0.3 & -0.2 \\ \hline 0.3 & -0.2 & 0.3 \end{array}\right] = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix},$$

$$\tilde{A} = A + E = \left[\begin{array}{cc|c} 6.5 & -1.1 & 1.3 \\ 0.6 & 4.3 & -0.2 \\ \hline 0.3 & -0.2 & 1.3 \end{array}\right] = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}.$$

The Jordan form of A is $\begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\quad \sigma(A_{11}) = \{5, 5\}, \quad \sigma(A_{22}) = \{1\}.$

The eigenmatrix of A is $Z_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, which satisfies $AZ_1 = Z_1 A_{11}$.

Question 1: Compute $\| T^{-1} \| = \| \begin{bmatrix} 5 & 1 \\ -1 & 3 \end{bmatrix}^{-1} \|_\infty = \frac{3}{8}$, and $\|A_{12}\|_\infty = 1$, $\|E_{12}\|_\infty = 0.3$, $\| E_{21} \|_\infty = 0.5$, $\| E_{11} \|_\infty = 0.7$, $\| E_{22} \|_\infty = 0.3$, to make sure the conditions in Theorem 5.2.10, which are $\tilde{\sigma} = 0.6, \quad \tilde{\eta} = 1.3, \quad \tilde{\gamma} = 0.5$. Then check $4\tilde{\eta}\tilde{\sigma}^2\tilde{\gamma} = \frac{117}{125} < 1$, i.e., (5.2.35) is satisfied.

Question 2: From theorem 5.2.9, take $A_0 = \tilde{A}$. For all $m = 0, 1, 2, \cdots$, we solve

$$P_m A_{11}^{(m)} - A_{22}^{(m)} P_m = A_{21}^{(m)},$$

and get

$$P_m = A_{21}^{(m)} S_m, \text{ where } S_m = \left(A_{11}^{(m)} - \begin{bmatrix} A_{22}^{(m)} & 0 \\ 0 & A_{22}^{(m)} \end{bmatrix} \right)^{-T}.$$

Compute $A_{11}^{(m+1)} = A_{11}^{(m)} + A_{12} P_m$, $\quad A_{22}^{(m+1)} = A_{22}^{(m)} - P_m A_{12}$ and $A_{21}^{(m+1)} = -P_m A_{12} P_m$. And then go back to solve P_{m+1}. Then

$$P^{(m)} = \sum_{i=0}^{m} P_i.$$

Compute $\| A_{21}^{(m)} \|_\infty$=? when $m = 0, 1, 2, 3.$ and $\| P_m \|_\infty$=? when $m = 0, 1, 2, 3.$

Compute $A_4 = \begin{bmatrix} I & 0 \\ -P^{(3)} & I \end{bmatrix} \tilde{A} \begin{bmatrix} I & 0 \\ P^{(3)} & I \end{bmatrix} \approx$?.

Compute $\tilde{Z}_1 \approx \tilde{Z}_1^{(3)} = \begin{bmatrix} I \\ P^{(3)} \end{bmatrix}$ =?.

Compute $\tilde{A}' = A_{11}^{(4)}$ =?.

5.3 Power Iterations

Given $A \in \mathbb{C}^{n \times n}$ and a unitary $U_0 \in \mathbb{C}^{n \times n}$. Consider the following iteration:

$$T_0 = U_0^* A U_0, \text{ for k=1,2,3, } \cdots \tag{5.3.1}$$

where $T_{k-1} = U_k R_k$ is the QR factorization of T_{k-1} and set $T_k = R_k U_k$. Since

$$T_k = (U_0 U_1 \cdots U_k)^* A (U_0 U_1 \cdots U_k), \tag{5.3.2}$$

it is obvious that each T_k is unitary similar to A.

Is (5.3.2) always "converges" to a Schur decomposition of A ?

Iteration (5.3.1) is called the **QR iterations**. (See Section 5.4)

5.3.1 Power Method

Let A be a diagonalizable matrix,

$$Ax_i = \lambda_i x_i, \quad i = 1, \cdots, n \tag{5.3.3}$$

with

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \tag{5.3.4}$$

and let $u_0 \neq 0$ be a given vector. From the expansion

$$u_0 = \sum_{i=1}^{n} \alpha_i x_i \tag{5.3.5}$$

follows that

$$A^s u_0 = \sum_{i=1}^{n} \alpha_i \lambda_i^s x_i = \lambda_1^s \{\alpha_1 x_1 + \sum_{i=2}^{n} \alpha_i (\frac{\lambda_i}{\lambda_1})^s x_i\}. \tag{5.3.6}$$

Thus the sequence of $\lambda_1^{-s} A^s u_0$ converges to a multiplicity of x_1. We consider two possibilities of normalization :

(A) $\|\quad\|$ - a given vector norm:

$$\begin{array}{ll} \text{For} & i = 0, 1, 2, \ldots, \\ & v_{i+1} = Au_i \\ & k_{i+1} = \|v_{i+1}\| \\ & u_{i+1} = v_{i+1}/k_{i+1} \text{ with initial } u_0 \\ \text{End} & \end{array} \tag{5.3.7}$$

Theorem 5.3.1 *Under the assumption* (5.3.4) *and* $\alpha_1 \neq 0$ *in* (5.3.5) *holds for the sequence defined by* (5.3.7)

$$\begin{aligned} \lim_{i\to\infty} k_i &= |\lambda_1| \\ \lim_{i\to\infty} \varepsilon^i u_i &= \frac{x_1}{\|x_1\|}\frac{\alpha_1}{|\alpha_1|}, \ \textit{where } \varepsilon = \frac{|\lambda_1|}{\lambda_1} \end{aligned}$$

Proof: It is obvious that

$$u_s = A^s u_0 / \|A^s u_0\| \ , k_s = \|A^s u_0\| / \|A^{s-1} u_0\|. \tag{5.3.8}$$

This follows from ${\lambda_1}^{-s} A^s u_0 \to \alpha_1 x_1$ that

$$\begin{aligned} |\lambda_1|^{-s} \|A^s u_0\| &\to |\alpha_1| \|x_1\|, \\ |\lambda_1|^{-s+1} \|A^{s-1} u_0\| &\to |\alpha_1| \|x_1\|, \end{aligned}$$

and then

$$|\lambda_1|^{-1} \|A^s u_0\| / \|A^{s-1} u_0\| = |\lambda_1|^{-1} k_s \longrightarrow 1.$$

From (5.3.6) follows now for $s \to \infty$

$$\varepsilon^s u_s = \varepsilon^s \frac{A^s u_0}{\|A^s u_0\|} = \frac{\alpha_1 x_1 + \sum \cdots}{\|\alpha_1 x_1 + \sum \cdots\|} \to \frac{\alpha_1 x_1}{\|\alpha_1 x_1\|} = \frac{x_1}{\|x_1\|}\frac{\alpha_1}{|\alpha_1|}. \tag{5.3.9}$$

■

(B) Let l be a linear functional:

Consider

$$\begin{aligned} &\text{For } \ i = 0, 1, 2, \ldots, \\ &\qquad v_{i+1} = Au_i, \\ &\qquad k_{i+1} = l(v_{i+1}) \quad \text{e.g. } e_n(v_{i+1}), e_1(v_{i+1}), \\ &\qquad u_{i+1} = v_{i+1}/k_{i+1} \quad \text{with initial } u_0. \\ &\text{End} \end{aligned} \tag{5.3.10}$$

Then it holds

Theorem 5.3.2 *Under the assumption of theorem 5.3.1, consider the method defined by (5.3.10) and suppose that $l(v_i) \neq 0$, for $i = 1, 2, \cdots$, and $l(x_1) \neq 0$. Then holds*

$$\lim_{i\to\infty} k_i = \lambda_1 \qquad \textit{and} \qquad \lim_{i\to\infty} u_i = \frac{x_1}{l(x_1)}.$$

Proof: As above we show that

$$u_i = A^i u_0 / l(A^i u_0) \ , k_i = l(A^i u_0)/l(A^{i-1} u_0).$$

From (5.3.6) we get for $s \to \infty$

$$\begin{aligned} {\lambda_1}^{-s} l(A^s u_0) &\to \alpha_1 l(x_1), \\ {\lambda_1}^{-s+1} l(A^{s-1} u_0) &\to \alpha_1 l(x_1), \end{aligned}$$

thus

$${\lambda_1}^{-1} k_s \to 1.$$

Similarly for $i \to \infty$,

$$u_i = \frac{A^i u_0}{l(A^i u_0)} = \frac{\alpha_1 x_1 + \sum_{j=2}^n \alpha_j (\frac{\lambda_j}{\lambda_1})^i x_j}{l(\alpha_1 x_1 + \sum \cdots)} \to \frac{\alpha_1 x_1}{\alpha_1 l(x_1)} \tag{5.3.11}$$

■

Remark 5.3.1 **(a)** *As linear functional l, a fix component k will always be chosen $l(x) = x_k$, k fix.*

(b) *The above argument also holds, if λ is a multiple eigenvalue.*

(c) *The iteration (5.3.10) follows*

$$\begin{aligned} k_s &= \frac{l(A^s u_0)}{l(A^{s-1} u_0)} = \lambda_1 \frac{\alpha_1 l(x_1) + \sum_{j=2}^n \alpha_j (\frac{\lambda_j}{\lambda_1})^s l(x_j)}{\alpha_1 l(x_1) + \sum_{j=2}^n \alpha_j (\frac{\lambda_j}{\lambda_1})^{s-1} l(x_j)} \\ &= \lambda_1 + O(\mid \frac{\lambda_2}{\lambda_1} \mid^{s-1}). \end{aligned} \tag{5.3.12}$$

That is the convergence depends on $\mid \frac{\lambda_2}{\lambda_1} \mid$. In the case $|\frac{\lambda_2}{\lambda_1}| = 1$ the iteration does not converge. Sometimes one can make the number $|\frac{\lambda_2}{\lambda_1}|$ small if we replace A with $A + \alpha I$, then the eigenvalue λ_i of A are transformed into $\lambda_i + \alpha$ and the convergence will be described by $(max_{i=1} |\frac{\lambda_i + \alpha}{\lambda_1 + \alpha}|)^s$. But this correction is not remarkable. The more useful method is use the inverse iteration . (See later !)

Now consider the case : A real and

$$\lambda_1 = \bar{\lambda}_2, \;\; |\lambda_1| > |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{5.3.13}$$

We can choose $x_2 = \bar{x}_1$ such that

$$Ax_1 = \lambda_1 x_1, \;\; A\bar{x}_1 = \bar{\lambda}_1 \bar{x}_1 = \lambda_2 \bar{x}_1.$$

Let u_0 be real and let

$$u_0 = \alpha_1 x_1 + \bar{\alpha}_1 \bar{x}_1 + \sum_{i\geq 3}^{n} \alpha_i x_i, \;\; \lambda_1 = \gamma e^{i\beta}, \;\; \alpha_1 = \rho e^{i\alpha}.$$

Then from (5.3.6) and (5.3.10) we have

$$\begin{aligned} A^s u_0 &= \alpha_1 \lambda_1^s x_1 + \bar{\alpha_1} \bar{\lambda_1}^s \bar{x_1} + \sum_{i\geq 3}^{n} \alpha_i \lambda_i^s x_i \\ &= \gamma^s \{\rho e^{i(\alpha+s\beta)} x_1 + \rho e^{-i(\alpha+s\beta)} \bar{x_1} + \sum_{i\geq 3}^{n} \alpha_i (\frac{\lambda_i}{\gamma})^s x_i\}. \end{aligned}$$

It happens oscillation without convergence!

Let

$$h(\lambda) = (\lambda - \lambda_1)(\lambda - \bar{\lambda}_1) = \lambda^2 - p\lambda - q, \;\; p = \lambda_1 + \bar{\lambda}_1 \;\; q = -\lambda_1 \bar{\lambda}_1.$$

Then

$$(A^{s+2} - pA^{s+1} - qA^s)u_0 = \alpha_1 {\lambda_1}^s \underbrace{h(\lambda_1)}_{=0} x_1 + \bar{\alpha}_1 \bar{\lambda}_1^s \underbrace{h(\bar{\lambda}_1)}_{=0} \bar{x}_1 + \sum_{i=3}^{n} \alpha_i h(\lambda_i) {\lambda_i}^s x_i.$$

Together with

$$l(A^s u_0) = r^s \{\rho e^{i(\alpha+s\beta)} l(x_1) + \rho e^{-i(\alpha+s\beta)} l(\bar{x}_1) + \sum_{i=3}^{n} \alpha_i (\frac{\lambda_i}{\gamma})^s l(x_i)\}$$

follows

$$k_{s+2}k_{s+1}u_{s+2} - pk_{s+1}u_{s+1} - qu_s = \frac{(A^{s+2} - pA^{s+1} - qA^s)u_0}{l(A^s u_0)} \to 0.$$

In this limit case u_{s+2}, u_{s+1} and u_s are linearly dependent. For fix s we determine p_s and q_s such that

$$\|k_{s+2}k_{s+1}u_{s+2} - p_s k_{s+1}u_{s+1} - q_s u_s\|_2 = \min!$$

We have to project the lot of $k_{s+2}k_{s+1}u_{s+2}$ on the plane determined by $k_{s+1}u_{s+1}$ and u_s, this leads

$$k_{s+2}k_{s+1}u_{s+2} - pk_{s+1}u_{s+1} - qu_s \perp u_{s+i}, \;\; i = 0, 1$$

or

$$\begin{pmatrix} u_{s+1}^T u_{s+1} & u_{s+1}^T u_s \\ u_s^T u_{s+1} & u_s^T u_s \end{pmatrix} \begin{pmatrix} p_s k_{s+1} \\ q_s \end{pmatrix} = k_{s+1}k_{s+2} \begin{pmatrix} u_{s+1}^T u_{s+2} \\ u_s^T u_{s+2} \end{pmatrix}. \tag{5.3.14}$$

We can show that $p_s \to p$, $q_s \to q$.

5.3.2 Inverse Power Iteration

Let α be an approximate eigenvalue of λ_1, i.e., $\alpha \approx \lambda_1$, then $(\alpha I - A)^{-1}$ has eigenvalues $\frac{1}{\alpha-\lambda_1}, \frac{1}{\alpha-\lambda_2}, \cdots, \frac{1}{\alpha-\lambda_n}$. Substitute A by $(\alpha I - A)^{-1}$, then the convergence is determined by $\max_{i\neq 1} |\frac{\alpha-\lambda_1}{\alpha-\lambda_i}|$.

Consider

$$\begin{array}{ll} \text{For} & i = 0, 1, 2, \ldots, \\ & v_{i+1} = (\alpha I - A)^{-1} u_i, \\ & k_{i+1} = l(v_{i+1}), \\ & u_{i+1} = v_{i+1}/k_{i+1} \quad \text{with initial vector } u_0, \\ \text{End} & \end{array} \tag{5.3.15}$$

Let A and u_0 be given and satisfy (5.3.3) and (5.3.5) respectively. Then we have the following theorem.

Theorem 5.3.3 *If* $| \alpha - \lambda_1 | < |\alpha - \lambda_i|$, *for* $i \neq 1$ *and suppose that* $\alpha_1 \neq 0$, $l(x_1) \neq 0$, *and* $l(v_i) \neq 0$ *for all* i *in* (5.3.15) *then holds*

$$\begin{aligned} &\lim_{i\to\infty} k_i = \frac{1}{\alpha - \lambda_1}, \quad (\lambda_1 \approx \alpha - \frac{1}{k_i}) \\ &\lim_{i\to\infty} u_i = \frac{x_1}{l(x_1)}. \end{aligned} \tag{5.3.16}$$

Variant I: (5.3.15) with constant α.
Variant II: Updating α.

$$\begin{array}{ll} \text{Given} & \alpha_{(0)} = \alpha \text{ and } u_0. \\ \text{For} & i = 0, 1, 2, \ldots, \\ & v_{i+1} = (\alpha_{(i)} I - A)^{-1} u_i, \\ & k_{i+1} = l(v_{i+1}), \\ & u_{i+1} = \frac{v_{i+1}}{k_{i+1}} \text{ and } \alpha_{(i+1)} = \alpha_{(i)} - \frac{1}{k_{i+1}}. \\ \text{End} & \end{array} \tag{5.3.17}$$

Show that: The method (5.3.17) is quadratic convergence.
Let $\alpha_{(i)} \approx \lambda_1$, $u_i \approx x_1$, and $l(x_1) = 1$. The remaining components of x_1 are smaller than 1 (Here $l(z) = z_1$). Let

$$u_m = (1 + \varepsilon_1^{(m)}) x_1 + \sum_{j=2}^{n} \varepsilon_j^{(m)} x_j, \quad \tilde{\varepsilon}^{(m)} = |\alpha^{(m)} - \lambda_1| \tag{5.3.18}$$

and

$$\delta_m = \max(|\varepsilon_1^{(m)}|, \cdots, |\varepsilon_n^{(m)}|, \tilde{\varepsilon}^{(m)}). \tag{5.3.19}$$

Claim: There exist a constant M independent on m with

$$\delta_{m+1} \leq M \delta_m^2. \tag{5.3.20}$$

Let $\varepsilon_i^{(m)} \approx \varepsilon_i$. Then we have

$$v_{m+1} = k_{m+1} u_{m+1} = \frac{1 + \varepsilon_1}{\alpha_{(m)} - \lambda_1} x_1 + \sum_{j=2}^{n} \frac{\varepsilon_j}{\alpha_{(m)} - \lambda_j} x_j \tag{5.3.21}$$

and

$$\begin{aligned}\alpha_{(m+1)} &= \alpha_{(m)} - \frac{1}{k_{m+1}} = \alpha_{(m)} - (\alpha_m - \lambda_1)[(1+\varepsilon_1) + \sum_{j=2}^{n} \frac{\varepsilon_j(\alpha_{(m)} - \lambda_1)}{\alpha_{(m)} - \lambda_j} x_{j,1}]^{-1}. \\ &= \alpha_{(m)} - (\alpha_{(m)} - \lambda_1)(1 + O(\delta_m)) = \lambda_1 + \delta_m O(\delta_m). \end{aligned} \tag{5.3.22}$$

From (5.3.21),

$$\begin{aligned} u_{m+1} &= \frac{k_{m+1}u_{m+1}}{k_{m+1}} \\ &= [(1+\varepsilon_1)x_1 + \sum_{j\geq 2} \frac{\varepsilon_j(\alpha_{(m)} - \lambda_1)}{\alpha_{(m)} - \lambda_j} x_j][1 + \varepsilon_1 + \sum_{j\geq 2} \frac{\varepsilon_j(\alpha_{(m)} - \lambda_1)}{\alpha_{(m)} - \lambda_j} x_{j,1}]^{-1} \\ &= [x_1 + \sum_{j\geq 2} \frac{\varepsilon_j(\alpha_{(m)} - \lambda_1)}{(1+\varepsilon_1)(\alpha_{(m)} - \lambda_j)} x_j] \underbrace{[1 + \sum_{j\geq 2} \frac{\varepsilon_j(\alpha_{(m)} - \lambda_1)}{(1+\varepsilon_1)(\alpha_{(m)} - \lambda_j)} x_{j,1}]^{-1}}_{1+O(\delta_m^2)}. \\ &= (1 + \varepsilon_1^{(m+1)})x_1 + \sum_{j\geq 2} \varepsilon_1^{(m+1)} x_j \end{aligned}$$

with $\varepsilon_i^{(m+1)} = O(\delta_m^2)$. This implies $\delta_{m+1} \leq M\delta_m^2$.

5.3.3 Connection with Newton-method

Consider the nonlinear equations

$$\begin{aligned} &Au - \lambda u = 0, \\ &l^T u = 1, \end{aligned}$$

for $n+1$ unknowns u and λ. Let

$$F\begin{pmatrix} u \\ \lambda \end{pmatrix} := \begin{pmatrix} Au - \lambda u \\ l^T u - 1 \end{pmatrix} = 0. \tag{5.3.23}$$

Newton method for (5.3.23):

$$\begin{pmatrix} u_{i+1} \\ \lambda_{i+1} \end{pmatrix} \equiv \begin{pmatrix} u \\ \lambda \end{pmatrix}_{(i+1)} = \begin{pmatrix} u \\ \lambda \end{pmatrix}_{(i)} - F'\begin{pmatrix} u \\ \lambda \end{pmatrix}_{(i)}^{-1} \begin{pmatrix} F\begin{pmatrix} u \\ \lambda \end{pmatrix}_{(i)} \end{pmatrix},$$

where

$$F'\begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} A - \lambda I & -u \\ l^T & 0 \end{pmatrix}.$$

Multiplying with $F'\binom{u}{\lambda}_{(i)}$ and write the first n equations and the last equation separately and simplify

$$(A - \lambda_i I)u_{i+1} = (\lambda_{i+1} - \lambda_i)u_i, \quad l^T u_{i+1} = 1. \tag{5.3.24}$$

We see that (5.3.24) identifies with (5.3.17) and is also quadratic convergence.

Variant III: A is real symmetric and α_{m+1} = Rayleigh Quotient

$$\begin{array}{l}\text{Give } u_0 \text{ with } \|u_0\|_2 = 1 \text{ and } \alpha_0 = u_0^T A u_0,\\ \text{For } m = 0, 1, 2, \ldots,\\ \qquad v_{m+1} = (\alpha_m I - A)^{-1} u_m,\\ \qquad u_{m+1} = \frac{v_{m+1}}{\|v_{m+1}\|_2},\\ \qquad \alpha_{m+1} = u_{m+1}^T A u_{m+1}.\\ \text{End}\end{array} \tag{5.3.25}$$

Claim: The iteration (5.3.25) is cubic convergence.

The eigenvectors x_i of A form an orthonormal system

$$x_i^T x_j = \delta_{ij}. \tag{5.3.26}$$

As above, let $\varepsilon_i^{(m)}$ and δ_m be defined in (5.3.18) and (5.3.19). From (5.3.26) follows $(\varepsilon_i = \varepsilon_i^{(m)})$:

$$\|u_m\|_2^2 = 1 = (1+\varepsilon_1)^2 + \sum_{j\geq 2} \varepsilon_j^2 = 1 + 2\varepsilon_1 + \sum_{j=1}^{n} \varepsilon_j^2.$$

So $\varepsilon_1 \leq \frac{n}{2}\delta_m^2 = O(\delta_m^2)$. That is

$$\begin{aligned} \alpha_{(m)} &= u_m^T A u_m = \lambda_1 (1+\varepsilon_1)^2 + \sum_{j\geq 2} \lambda_j \varepsilon_j^2 \\ &= \lambda_1 + 2\varepsilon_1 \lambda_1 + \sum_{j=1}^{n} \lambda_j \varepsilon_j^2 = \lambda_1 + O(\delta_m^2). \end{aligned} \tag{5.3.27}$$

Thus $\tilde{\varepsilon}^{(m)} = O(\delta_m^2)$. On the other hand,

$$\begin{aligned} v_{m+1}^T v_{m+1} &= \frac{(1+\varepsilon_1)^2}{(\alpha_{(m)} - \lambda_1)^2} + \sum_{j\geq 2} \frac{\varepsilon_j^2}{(\alpha_{(m)} - \lambda_j)^2} \\ &= |\frac{1+\varepsilon_1}{\alpha_{(m)} - \lambda_1}|^2 \{ \underbrace{1 + \sum_{j\geq 2} \frac{\varepsilon_j^2 \tilde{\varepsilon}^{(m)2}}{(1+\varepsilon_1)^2 (\alpha_{(m)} - \lambda_j)^2}}_{1+O(\delta_m^6)} \}. \end{aligned}$$

Therefore

$$\begin{aligned} u_{m+1} &= (\frac{1+\varepsilon_1}{\alpha_{(m)} - \lambda_1} x_1 + \sum_{j\geq 2} \frac{\varepsilon_j}{\alpha_{(m)} - \lambda_j} x_j)(1 + O(\delta_m^6))(\frac{\alpha_{(m)} - \lambda_1}{1+\varepsilon_1}) \\ &= [x_1 + \sum_{j\geq 2} \frac{\varepsilon_j \tilde{\varepsilon}^{(m)}}{(1+\varepsilon_1)(\alpha_{(m)} - \lambda_j)} x_j](1 + O(\delta_m^6)) \\ &= (1 + \varepsilon_1^{(m+1)}) x_1 + \sum_{j\geq 2} \varepsilon_j^{(m+1)} x_j \end{aligned}$$

with $|\ \varepsilon_j^{(m+1)}\ | \leq M\delta_m^3$ $(j = 1, \cdots, n)$. As in (5.3.27) we have

$$|\alpha_{(m+1)} - \lambda_1| = O(\delta_{m+1}^2) = O(\delta_m^6).$$

■

5.3.4 Orthogonal Iteration

Given $Q_0 \in \mathbb{C}^{n\times p}$ with orthogonal columns and $1 \le p < n$.

$$\begin{array}{l} \text{For } k = 1, 2, \cdots \\ \qquad Z_k = AQ_{k-1}, \\ \qquad Q_k R_k = Z_k, \ \text{(QR decomposition)} \\ \text{End} \end{array} \tag{5.3.28}$$

Note that if $p = 1$ this is just the power method. Suppose that

$$Q^* AQ = T = diag(\lambda_i) + N, \ \ |\lambda_1| \ge \cdots \ge |\lambda_n| \tag{5.3.29}$$

is a Schur decomposition of A and partition Q, T and N as follows

$$Q = [\overbrace{Q_\alpha}^{p}, \overbrace{Q_\beta}^{n-p}], T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}, \quad N = \begin{pmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{pmatrix} \tag{5.3.30}$$

If $|\lambda_p| > |\lambda_{p+1}|$ we define $D_p(A) = \mathcal{R}(Q_\alpha) \equiv \text{Range}(Q_\beta)$ is a **dominant** invariant subspace . It is the unique subspace associated with associated with $\lambda_1, \cdots, \lambda_p$. The following theorem (without proof see Golub/Vanloan p.215) shows that the subspace $\mathcal{R}(Q_k)$ generated by (5.3.28) converges to $D_p(A)$ at a rate proportional to $|\frac{\lambda_{p+1}}{\lambda_p}|^k$ under reasonable assumptions.

Theorem 5.3.4 *Let the Schur form of A be given by* (5.3.29) *and* (5.3.30). *Assume that $|\lambda_p| > |\lambda_{p+1}|$ and that $\theta \ge 0$ satisfies*

$$(1+\theta)|\lambda_p| > \|N\|_F.$$

If $Q_0 \in \mathbb{C}^{n\times p}$ with $Q_0^* Q_0 = I_p$ and $d = dist[D_p(A^*), \mathcal{R}(Q_0)] < 1$, then Q_k generated by (5.3.28) *satisfy*

$$dist[D_p(A), \mathcal{R}(Q_k)] \le \frac{(1+\theta)^{n-2}}{\sqrt{1-d^2}}[1 + \frac{\|T_{12}\|_F}{sep(T_{11}, T_{12})}][\frac{|\lambda_{p+1}| + \|N\|_F/(1+\theta)}{|\lambda_p| - \|N\|_F/(1+\theta)}]^k.$$

When θ is chosen large enough then the theorem essentially shows that

$$\text{dist}[D_p(A), R(Q_k)] \le c|\lambda_{p+1}/\lambda_p|^k,$$

where c depends on $sep(T_{11}, T_{12})$ and $\|N\|_F$. Needless to say, the convergence can be very slow if the gap between $|\lambda_p|$ and $|\lambda_{p+1}|$ is not sufficiently wide. To prove this theorem we need to prove the following two lemmas 5.3.1 and 5.3.3.

Lemma 5.3.1 *Let* $T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$ *and define the linear operator* $\varphi : \mathbb{C}^{p\times q} \to \mathbb{C}^{p\times q}$ *by*

$$\varphi(X) = T_{11}X - XT_{22}.$$

Then φ is nonsingular $\iff \sigma(T_{11}) \cap \sigma(T_{22}) = \phi$. if φ is nonsingular and $\varphi(Z) = -T_{12}$, then $Y^{-1}TY = diag(T_{11}, T_{22})$, where $Y = \begin{bmatrix} I_p & Z \\ 0 & I_q \end{bmatrix}$.

Proof: "$\Longleftarrow$": Suppose $\varphi(X) = 0$ for $X \neq 0$ and

$$U^*XV = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma_r = diag(\sigma_i), \quad r = rank(X).$$

Substituting into $T_{11}X = XT_{22}$ gives

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where

$$U^*T_{11}U = (A_{ij}) \text{ and } V^*T_{22}V = (B_{ij}).$$

By comparing blocks, we see that $A_{21} = B_{12} = 0$ and $\sigma(A_{11}) = \sigma(B_{11})$. Conversely, $\phi \neq \sigma(A_{11}) = \sigma(B_{11}) \subset \sigma(T_{11}) \cap \sigma(T_{22})$.

"$\Longrightarrow$": If $\lambda \in \sigma(T_{11}) \cap \sigma(T_{22})$, then there are $x \neq 0$, $y \neq 0$ satisfy $T_{11}x = \lambda x$ and $y^*T_{22} = \lambda y^*$. This implies $\varphi(xy^*) = 0$.

Finally, if φ nonsingular, then Z exists and

$$Y^{-1}TY = \begin{bmatrix} T_{11} & T_{11}Z - ZT_{22} + T_{12} \\ 0 & T_{22} \end{bmatrix} = \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix}.$$

■

{ Another proof }

For $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times m}$ define the Kronecker product of A and B by

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mm}B \end{bmatrix} = [a_{ij}B]_{i,j=1}^{m} \in \mathbb{C}^{mn\times mn}.$$

Let $C = [c_1, \cdots, c_n] \in \mathbb{C}^{m\times m}$. Define

$$\text{vec}(C) = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{C}^{mn\times 1}.$$

Consider the linear matrix equation

$$AX - XB = C. \tag{5.3.31}$$

Lemma 5.3.2 *vec(AX − XB) = (I ⊗ A − B^T ⊗ I)vec(X).*

Proof:

$$(AX)_j = AX_j \rightarrow \text{vec}(AX) = (I \otimes A)\text{vec}(X),$$

$$(XB)_j = \sum_{k=1}^{n} b_{kj}X_k = [b_{1j}I, \cdots, b_{nj}I]\text{vec}(X).$$

By linearity of vec we have

$$\text{vec}(AX - XB) = \text{vec}(AX) - \text{vec}(XB) = [(I \otimes A) - (B^T \otimes I)]\text{vec}(X).$$

■

Let $G = [(I \otimes A - B^T \otimes I)]$, $X = vec(X)$, $r = vec(C)$. Then the equation (5.3.31) is equivalent to $Gx = r$ and the equation (5.3.31) has a unique solution $\Longleftrightarrow \sigma(A) \cap \sigma(B) = \phi$. There are unitary Q_1, Z_1 such that

$$Q_1^* A Q_1 = A_1 = \begin{bmatrix} r_1 & * & * \\ 0 & \ddots & * \\ 0 & 0 & r_m \end{bmatrix}, \quad Z_1^* B Z_1 = B_1 = \begin{bmatrix} s_1 & 0 & 0 \\ * & \ddots & 0 \\ * & * & s_m \end{bmatrix}.$$

(5.3.31) becomes

$$\begin{aligned} & Q_1^* A Q_1 Q_1^* X Z_1 - Q_1^* X Z_1 Z_1^* B Z_1 = Q_1^* C Z_1 \equiv C_1 \\ \Longleftrightarrow\ & A_1 X_1 - X_1 B_1 = C_1, \text{ where } X_1 = Q_1^* X Z_1 \\ \Longleftrightarrow\ & G_1 x_1 = r_1, \end{aligned}$$

where $G_1 = [I \otimes A_1 - B_1 \otimes I]$ and $x_1 = \text{vec}(X_1)$, $r_1 = \text{vec}(C_1)$. Also

$$\det(G_1) = \prod_{\substack{1 \le i \le m \\ 1 \le j \le n}} (r_i - r_j).$$

Hence we have $\sigma(A) \cap \sigma(B) = \phi \Longleftrightarrow (r_j - s_j) \neq 0 \ \ (i = 1, \cdots m, j = 1, \cdots n) \Longleftrightarrow \det(G_1) \neq 0 \Longleftrightarrow G_1 x_1 = r_1$, has a unique solution. $\Longleftrightarrow$ the equation (5.3.31) has a unique solution X. ■

Exercise:

(a) Consider the linear matrix equation $AXB - CXD = R$ where $A, C \in \mathbb{C}^{m \times m}$, $B, D \in \mathbb{C}^{n \times n}$ and $X, R \in \mathbb{C}^{m \times n}$. The equation has a unique solution $\Longleftrightarrow \sigma(A, C) \cap \sigma(B, D) = \phi$.

(b) Consider $\begin{cases} AX - YB = R, \\ CX - YD = S, \end{cases}$ where $A, B, C, D, X, Y, R, S \in \mathbb{C}^{m \times n}$. The equation has a unique solution $(X, Y) \Longleftrightarrow \sigma(A, C) \cap \sigma(B, D) = \phi$.

Lemma 5.3.3 *Let $Q^*AQ = T = D + N$ (Schur decomposition). D is diagonal and N is strictly upper triangular. Let $\lambda = \max\{|\eta| : \det(A - \eta I) = 0\}$ and $\mu = min\{|\eta| : \det(A - \eta I)\} = 0$. If $\theta \ge 0$, then*

$$\|A^k\|_2 \le (1 + \theta)^{n-1} [|\lambda| + \frac{\|N\|_F}{1 + \theta}]^k, \quad k \ge 0. \tag{5.3.32}$$

If A is nonsingular and $\theta \ge 0$ satisfies $(1 + \theta)|\mu| > \|N\|_F$, then

$$\|A^{-k}\|_2 \le (1 + \theta)^{n-1} [\frac{1}{|\mu| - \|N\|_F/(1 + \theta)}]^k, \quad k \ge 0. \tag{5.3.33}$$

Proof: For $\theta \geq 0$, define $\triangle = \text{diag}(1, 1+\theta, (1+\theta)^2, \cdots, (1+\theta)^{n-1})$ and $\kappa_2(\triangle) = (1+\theta)^{n-1}$. But $\|\triangle N \triangle_F^{-1}\| \leq \|N\|_F/(1+\theta)$, thus

$$\begin{aligned} \|A^k\|_2 &= \|T^k\|_2 = \|\triangle^{-1}(D + \triangle N \triangle^{-1})^k \triangle\|_2 \\ &\leq \kappa_2(\triangle)[\|D\|_2 + \|\triangle N \triangle^{-1}\|_2]^k \\ &\leq (1+\theta)^{n-1}[|\lambda| + \|N\|_F/(1+\theta)]^k. \end{aligned}$$

On the other hand, if A is nonsingular and $(1+\theta)|\mu| > \|N\|_F$, then

$$\|\triangle D^{-1} N \triangle^{-1}\|_2 < 1$$

and thus,

$$\begin{aligned} \|A^{-k}\|_2 &= \|T^{-k}\|_2 = \|\triangle^{-1}(I + \triangle D^{-1} N \triangle^{-1})^{-1} D^{-1}]^k \triangle\|_2 \\ &\leq \kappa_2(\triangle)[\|D^{-1}\|_2/[1 - \|\triangle D^{-1} N \triangle^{-1}\|_2]^k \\ &\leq (1+\theta)^{n-1}[\frac{1}{|\mu| - \|N\|_F/(1+\theta)}]^k \end{aligned}$$

■

{ proof of Theorem 5.3.4: } By induction $A^k Q_0 = Q_k(R_k \cdots R_1)$. By substituting (5.3.29), (5.3.30) into this equality we get

$$\begin{bmatrix} V_0 \\ W_0 \end{bmatrix} = \begin{bmatrix} V_k \\ W_k \end{bmatrix} (R_k, \cdots, R_1),$$

where $V_k = Q_\alpha^* Q_k$ and $W_k = Q_\beta^* Q_k$. Using Lemma 5.3.1 there is an $X \in C^{p \times (n-p)}$ such that

$$\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix}.$$

Moreover since $\text{sep}(T_{11}, T_{22}) =$ the smallest singular value of $\phi(X) = T_{11}X - XT_{22}$. From $\phi(X) = -T_{12}$ follows

$$\|X\|_F \leq \|T_{12}\|_F / sep(T_{11}, T_{22}).$$

Thus

$$\begin{bmatrix} T_{11}^k & 0 \\ 0 & T_{22}^k \end{bmatrix} \begin{bmatrix} V_0 - XW_0 \\ W_0 \end{bmatrix} = \begin{bmatrix} V_k - XW_k \\ W_k \end{bmatrix} (R_k, \cdots, R_1).$$

Assume $V_0 - XW_0$ is nonsingular. Then

$$W_k = T_{22}^k W_0 (V_0 - XW_0)^{-1} T_{11}^{-k} (V_k - XW_k).$$

From Theorem 5.1.3 follows that

$$\text{dist}[D_p(A), \mathcal{R}(Q_k)] = \|Q_\beta^* Q_k\|_2 = \|W_k\|_2.$$

Then

$$\text{dist}[D_p(A), \mathcal{R}(Q_k)] \leq \|T_{22}^k\|_2 \|(V_0 - XW_0)^{-1}\|_2 \|T_{11}^{-k}\|_2 [1 + \|X\|_F]. \tag{5.3.34}$$

We prove $V_0 - XW_0$ is nonsingular. From $A^*Q = QT^*$ follows that

$$A^*(Q_\alpha - Q_\beta X^*) = (Q_\alpha - Q_\beta X^*)T_{11}^*,$$

which implies orthogonal column of $Z = (Q_\alpha - Q_\beta X^*)(I + XX^*)^{-\frac{1}{2}}$ are a basis of $D_p(A^*)$. Also

$$(V_0 - XW_0) = (I + XX^*)^{\frac{1}{2}} Z^* Q_0.$$

This implies

$$\sigma_p(V_0 - XW_0) \geq \sigma_p(Z^*Q_0) = \sigma_p(V_0 - XW_0) \geq \sigma_p(Z^*Q_0) = \sqrt{1 - d^2} > 0.$$

Hence $V_0 - XW_0$ is invertible and $\|(V_0 - XW_0)^{-1}\|_2 \leq \frac{1}{\sqrt{1-d^2}}$. By Lemma 5.3.1 we get

$$\|T_{22}^k\|_2 \leq (1+\theta)^{n-p-1}[|\lambda_{p+1}| + \|N\|_F/(1+\theta)]^k.$$

and

$$\|T_{11}^{-k}\|_2 \leq (1+\theta)^{p-1}/[|\lambda_p| - \|N\|_F/(1+\theta)]^k.$$

Substituting into (5.3.34) the theorem is proved. ■

5.4 QR-algorithm (QR-method, QR-iteration)

Theorem 5.4.1 (Schur Theorem) *There exists a unitary matrix U such that*

$$AU = UR,$$

where R is upper triangular.

Iteration method (from Vojerodin):

$$\begin{array}{l} \text{Set } \ U_0 = I, \\ \text{For } \ i = 0, 1, 2, \cdots \\ \qquad AU_i = U_{i+1}R_{i+1}, \quad \text{(an QR factorization of } AU_i\text{.)} \\ \text{End} \end{array} \tag{5.4.1}$$

If U_i converges to U, then for $i \to \infty$

$$R_{i+1} = U_{i+1}^* A U_i \to U^* A U.$$

We now define

$$Q_i = U_{i-1}^* U_i, \ \ A_{i+1} = U_i^* A U_i. \tag{5.4.2}$$

Then from (5.4.1) we have

$$A_i = U_{i-1}^* A U_{i-1} = U_{i-1}^* U_i R_i = Q_i R_i.$$

On the other hand from (5.4.1) substituting i by $i-1$ we get

$$R_i U_{i-1}^* = U_i^* A$$

and thus

$$R_iQ_i = R_iU_{i-1}^*U_i = U_i^*AU_i = A_{i+1}.$$

So (5.4.1) for $U_0 = I$ and $A_1 = A$ is equivalent to:

For $i = 1, 2, 3, \cdots$

$$A_i = Q_iR_i \quad \text{(QR factorization of } A_i\text{)}, \tag{5.4.3}$$

$$A_{i+1} = R_iQ_i. \tag{5.4.4}$$

End

Equations (5.4.3)-(5.4.4) describe the basic form of QR algorithm. We prove two important results. Let

$$P_i = Q_1Q_2\cdots Q_i, \quad S_i = R_iR_{i-1}\cdots R_1. \tag{5.4.5}$$

Then hold

$$A_{i+1} = P_i^*AP_i = S_iAS_i^{-1}, \quad i = 1, 2, \cdots \tag{5.4.6}$$

$$A^i = P_iS_i \quad i = 1, 2, \cdots. \tag{5.4.7}$$

(5.4.6) is evident. (5.4.7) can be proved by induction. For $i = 1, A_1 = Q_1R_1$, Suppose (5.4.7) holds for i. Then

$$\begin{aligned} A^{i+1} &= AP_iS_i = P_iA_{i+1}S_i \qquad \text{(from (5.4.6))} \\ &= P_iQ_{i+1}R_{i+1}S_i = P_{i+1}S_{i+1}. \end{aligned}$$

Theorem 5.4.2 *Let $A \in \mathbb{C}^{n\times n}$ with eigenvalues λ_i under the following assumptions:*

(a)

$$|\lambda_1| > |\lambda_2| > \cdots |\lambda_n| > 0; \tag{5.4.8}$$

(b) *The factorization*

$$A = X\Lambda X^{-1} \tag{5.4.9}$$

with $X^{-1} = Y$ and $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ holds. Here Y has an LR factorization.

Then QR algorithm converges. Furthermore

(a) $\lim_{i\to\infty} a_{jk}^{(i)} = 0$, *for $j > k$, where $A_i = (a_{jk}^{(i)})$;*

(b) $\lim_{i\to\infty} a_{kk}^{(i)} = \lambda_k$, *for $k = 1, \cdots, n$.*

Remark 5.4.1 *Assumption* (5.4.9) *is not essential for convergence of the QR algorithm. If the assumption is not satisfied, the QR algorithm still converges, only the eigenvalues on the diagonal no longer necessary appear ordered in absolute values, i.e.* (b) *is replaced by* (b') $\lim_{i\to\infty} a_{kk}^{(i)} = \lambda_{\pi(k)}$, *$k = 1, 2\cdots, n$, where π is a permutation of $\{1, 2, \cdots, n\}$. (See Wilkinson pp.519)*

Proof: { of Theorem 5.4.2 } Let $X = QR$ be the QR factorization of X with $r_{ii} > 0$ and $Y = LU$ be the LR factorization of Y with $\ell_{ii} = 1$. Since $A = X\Lambda X^{-1} = QR\Lambda R^{-1}Q^*$, we have

$$Q^*AQ = R\Lambda R^{-1} \tag{5.4.10}$$

is an upper-triangular matrix with diagonal elements λ_i ordered in absolute value as in (5.4.8). Now

$$A^s = X\Lambda^s X^{-1} = QR\Lambda^s LU = QR\Lambda^s L\Lambda^{-s}\Lambda^s U$$

and since

$$(\Lambda^s L\Lambda^{-s})_{ik} = \ell_{ik}(\frac{\lambda_i}{\lambda_k})^s = \begin{cases} 0, & i < k, \\ 1, & i = k, \\ \to 0, & i > k \ \text{ as } \ s \to \infty, \end{cases}$$

where $\Lambda^s L\Lambda^{-s} = I + E_s$, with $\lim_{s\to\infty} E_s = 0$. Therefore

$$A^s = QR(I + E_s)\Lambda^s U = Q(I + RE_sR^{-1})R\Lambda^s U = Q(I + F_s)R\Lambda^s U$$

with $\lim_{s\to\infty} F_s = 0$. From the conclusion of QR factorization the matrices Q and R ($r_{ii} > 0$) depend continuously on A ($A = QR$). But $I = I \cdot I$ is the QR factorization of I, therefore it holds for the QR factorization:

$$I + F_s = \tilde{Q}_s\tilde{R}_s.$$

Thus for $F_s \to 0$, we have $\lim_{s\to\infty} \tilde{Q}_s = I$ and $\lim_{s\to\infty} \tilde{R}_s = I$. From (5.4.7) we have

$$A^s = (Q\tilde{Q}_s)(\tilde{R}_sR\Lambda^s U) = P_sR_s.$$

So from the "uniqueness" of QR factorization there exists a unitary diagonal matrix D_s with

$$P_sD_s = Q\tilde{Q}_s \to Q.$$

Thus from (5.4.6) we have

$$D_i^*A_{i+1}D_i = D_i^*P_i^*AP_iD_i \to Q^*AQ = R\Lambda R^{-1}. \tag{5.4.11}$$

The assertions (a) and (b) are proved. ■

Remark 5.4.2 *One can show that* $\lim_{s\to\infty} Q_s = diag(\frac{\lambda_i}{|\lambda_i|})$. *That is in general* Q_s *does not converge to* I *and then* P_s *does not converge. Therefore* D_s *does not converge to* I *and* (5.4.11) *shows that the elements of* A_s *over the diagonal elements oscillate and only converge in absolute values.*

Let A be diagonalizable and the eigenvalues such that

$$|\lambda_1| = \cdots = |\lambda_{\nu_1}| > |\lambda_{\nu_1+1}| = \cdots = |\lambda_{\nu_2}| > \cdots = |\lambda_{\nu s}| \tag{5.4.12}$$

with $\nu_s = n$. We define a block partition of $n \times n$ matrix B in s^2 blocks $B_{k\ell}$ for $k, \ell = 1, 2, \cdots, s$

$$B = [B_{k\ell}]_{k,\ell=1}^s.$$

Theorem 5.4.3 (Wilkinson) *Let A be diagonalizable and satisfy* (5.4.12) *and* (5.4.9). *Then it holds for the blocks $A_{jk}^{(i)}$ of A_i that*

(a) $\lim_{i\to\infty} A_{jk}^{(i)} = 0, \quad j > k;$

(b) *The eigenvalues of $A_{kk}^{(i)}$ converges to the eigenvalues $\lambda_{\nu_{k-1}+1}, \cdots, \lambda_{\nu_k}$.*

Special case: If A is real and all the eigenvalues have different absolute value except conjugate eigenvalues. Then

$$A_i \to \begin{bmatrix} \times & \times & + & + & + & + & + \\ \times & \times & + & + & + & + & + \\ & & \times & \times & + & + & + \\ & & \times & \times & + & + & + \\ & & & & \times & + & + \\ & & & & & \times & \times \\ 0 & & & & & \times & \times \end{bmatrix}.$$

Theorem 5.4.4 *Let A be an upper Hessenberg matrix. Then the matrices Q_i and A_i in* (5.4.3) *and* (5.4.4) *are also upper Hessenberg matrices.*

Proof: It is obvious from $A_{i+1} = R_i A_i R_i^{-1}$ and $Q_i = A_i R_i^{-1}$. ■

5.4.1 The Practical QR Algorithm

In the following paragraph we will develop an useful QR algorithm for real matrix A. We will concentrate on developing the iteration

$$\begin{aligned} &\text{Compute orthogonal } Q_0 \text{ such that } H_0 = Q_0^T A Q_0 \text{ is upper Hessenberg.} \\ &\text{For } \ k = 1, 2, 3, \cdots \\ &\qquad \text{Compute QR factorization } \ H_k = Q_k R_k; \\ &\qquad \text{Set } H_{k+1} = R_k Q_k; \\ &\text{End} \end{aligned} \tag{5.4.13}$$

Here $A \in \mathbb{R}^{n\times n}$, $Q_i \in \mathbb{R}^{n\times n}$ is orthogonal and $R_i \in \mathbb{R}^{n\times n}$ is upper triangular.

Theorem 5.4.5 (Real Schur Decomposition) *If $A \in \mathbb{R}^{n\times n}$, then there exists an orthogonal $Q \in \mathbb{R}^{n\times n}$ such that*

$$Q^T A Q = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{21} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{mm} \end{pmatrix} \tag{5.4.14}$$

where each R_{ii} is either 1×1 or 2×2 matrix having complex conjugate eigenvalues.

Proof: Let k be the number of complex conjugate pair in $\sigma(A)$. We prove the theorem by induction on k. The theorem holds if $k = 0$. Now suppose that $k \geq 1$. If $\lambda = \gamma + i\mu \in \sigma(A)$ and $\mu \neq 0$, then there exists vectors y and $z \in \mathbb{R}^n (z \neq 0)$ such that

$$A(y + iz) = (\gamma + i\mu)(y + iz),$$

i.e.,

$$A[y, z] = [y, z] \begin{bmatrix} \gamma & \mu \\ -\mu & \gamma \end{bmatrix}.$$

The assumption that $\mu \neq 0$ implies that y and z span a two dimensional, real invariant subspace for A. It then follows that

$$U^T AU = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \text{ with } \sigma(T_{11}) = \{\lambda, \bar{\lambda}\}.$$

By induction, there exists an orthogonal $\tilde{U}$ so that $\tilde{U}^T T_{22} \tilde{U}$ has the require structure. The theorem follows by setting $Q = U diag(I_2, \tilde{U})$. ■

Algorithm 5.4.1 (Hessenberg QR step)

Input: Given the upper Hessenberg matrix $H \in \mathbb{R}^{n \times n}$;
  Compute QR factorization of H: $H = QR$ and overwrite H with $\bar{H} = RQ$;
  For $k = 1, \cdots, n-1$,
    Determine c_k and s_k with $c_k^2 + s_k^2 = 1$ such that

$$\begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix} \begin{bmatrix} h_{kk} \\ h_{k+1,k} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$$

    For $j = k, \cdots, n$,

$$\begin{bmatrix} h_{kj} \\ h_{k+1,j} \end{bmatrix} = \begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix} \begin{bmatrix} h_{kj} \\ h_{k+1,j} \end{bmatrix}.$$

    End;
  End;
  For $k = 1, \cdots, n-1$,
    For $i = 1, \cdots, k+1$,

$$[h_{ik}, h_{i,k+1}] \equiv [h_{ik}, h_{i,k+1}] \begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix}.$$

    End;
  End;

This algorithm requires $4n^2$ flops. Moreover, since $Q^T = J(n-1, n, \theta_{n-1}) \cdots J(1, 2, \theta_1)$ is lower Hessenberg $\bar{H} = QR$ is upper Hessenberg. Thus the QR iteration preserves Hessenberg structure.

We now describe how the Hessenberg decomposition $Q_0^T A Q_0 = H$ =upper Hessenberg to be computed.

Algorithm 5.4.2 (Householder Reduction to Hessenberg Form) *Given $A \in \mathbb{R}^{n \times n}$. The following algorithm overwrites A with $H = Q_0^T A Q_0$, where H is upper Hessenberg and $Q_0 = P_1 \cdots P_{n-2}$ is a product of Householder matrices.*

For $k = 1, \cdots, n-2$,
Determine a Householder matrix $\bar{P}_k$ *of order* $n-k$ *such that*

$$\bar{P}_k \begin{bmatrix} a_{k+1,k} \\ \vdots \\ \vdots \\ a_{n,k} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Compute $A \equiv P_k^T A P_k$ *where* $P_k = diag(I_k, \bar{P}_k)$.
End;

This algorithm requires $\frac{5}{3}n^3$ flops. Q_0 can be stored in factored form below the subdiagonal A. If Q_0 is explicitly formed, an additional $\frac{2}{3}n^3$ flops are required.

Theorem 5.4.6 (Implicit Q Theorem) *Suppose* $Q = [q_1, \cdots, q_n]$ *and* $V = [v_1, \cdots, v_n]$ *are orthogonal matrices with* $Q^T A Q = H$ *and* $V^T A V = G$ *are upper Hessenberg. Let* k *denote the smallest positive integer for which* $h_{k+1,k} = 0$ *with the convention that* $k = n$, *if* H *is unreduced. If* $v_1 = q_1$, *then* $v_i = \pm q_i$ *and* $|h_{i,i-1}| = |g_{i,i-1}|$, *for* $i = 2, \cdots, k$. *Moreover if* $k < n$ *then* $g_{k+1,k} = 0$.

Proof: Define $W = V^T Q = [w_1, \cdots, w_n]$ orthogonal, and observe $GW = WH$. For $i = 2, \cdots k$, we have

$$h_{i,i-1} w_i = G w_{i-1} - \sum_{j=1}^{i-1} h_{j,i-1} w_j$$

Since $w_1 = e_1$, it follows that $[w_1, \cdots, w_k]$ is upper triangular and thus $w_i = \pm e_i$ for $i = 2, \cdots, k$. Since $w_i = V^T q_i$ and $h_{i,i-1} = w_i^T G w_{i-1}$, it follows that $v_i = \pm q_i$ and $|h_{i,i-1}| = |g_{i,i-1}|$ for $i = 2, \cdots, k$. If $h_{k+1,k} = 0$, then ignoring signs we have

$$\begin{aligned} g_{k+1,k} &= e_{k+1}^T G e_k = e_{k+1}^T G W e_k = (e_{k+1}^T W)(H e_k) \\ &= e_{k+1}^T \sum_{i=1}^{k} h_{ik} W e_i = \sum_{i=1}^{k} h_{ik} e_{k+1}^T e_i = 0. \end{aligned}$$

■

Remark 5.4.3 *The gist of the implicit Q theorem is that if* $Q^T A Q = H$ *and* $Z^T A Z = G$ *are each unreduced upper Hessenberg matrices and* Q *and* Z *have the same first column, then* G *and* H *are "essentially equal" in the sense that* $G = D^{-1} H D$, *where* $D = diag(\pm 1, \cdots, \pm 1)$.

We now return to Hessenberg QR iteration in (5.4.13):

Give orthogonal Q_0 such that $H = Q_0^T A Q_0$ is upper Hessenberg.
For $k = 1, 2, 3, \cdots$
$H = QR$, (QR factorization)
$H := RQ$, (upper Hessenberg)
End

Without loss of generality we may assume that each Hessenberg matrix produced by (5.4.13) is unreduced. If not, then at some stage we have

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix} \text{ with } H_{11} \in \mathbb{R}^{p\times p} \quad (1 \le p < n).$$

The problem "decouples" into two small problems involving H_{11} and H_{22}. The term "deflation" is also used in this context, usually when $p = n-1$ or $n-2$. In practice, decoupling occurs whenever a subdiagonal entry in H is suitably small. For example in EISPACK if

$$|h_{p+1,p}| \le \text{eps}(|h_{p,p}| + |h_{p+1,p+1}|), \tag{5.4.15}$$

then $h_{p+1,p}$ is "declared" to be zero.

Now we will investigate how the convergence (5.4.13) can be accelerated by incorporating "shifts". Let $\mu \in \mathbb{R}$ and consider the iteration

$$\begin{array}{l} \text{Give orthogonal } Q_0 \text{ such that } H = Q_0^T A Q_0 \text{ is upper Hessenberg.} \\ \text{For } k = 1, 2, \cdots \\ \qquad H - \mu I = QR, \quad \text{(QR factorization)} \\ \qquad H = RQ + \mu I, \\ \text{End} \end{array} \tag{5.4.16}$$

The scale μ is refereed to a shift. Each matrix H in (5.4.16) is similar to A, since $RQ + \mu I = Q^T(QR + \mu I)Q = Q^T H Q$.

If we order the eigenvalues λ_i of A so that $|\lambda_1 - \mu| \ge \cdots \ge |\lambda_n - \mu|$, then Theorem 5.4.5 says that the p-th subdiagonal entry in H converges to zero with rate $|\frac{\lambda_{p+1}-\mu}{\lambda_p - \mu}|^k$. Of course if $\lambda_p = \lambda_{p+1}$ then there is no convergence at all. But if μ is much closer to λ_n than to the other eigenvalues, the convergence is required.

Theorem 5.4.7 *Let μ be an eigenvalues of an $n \times n$ unreduced Hessenberg matrix H. If $\bar{H} = RQ + \mu I$, where $(H - \mu I) = QR$ is the QR decomposition of $H - \mu I$, then $\bar{h}_{n,n-1} = 0$ and $\bar{h}_{nn} = \mu$.*

Proof: If H is unreduced, then so is the upper Hessenberg matrix $H - \mu I$. Since $Q^T(H - \mu I) = R$ is singular and since it can be shown that

$$|r_{ii}| \ge |h_{i+1,i}|, \quad i = 1, 2, \cdots, n-1, \tag{5.4.17}$$

it follows that $r_m = 0$. Consequently, the bottom row of $\bar{H}$ is equal to $(0, \cdots, 0, \quad \mu)$. ■

5.4.2 Single-shift QR-iteration

$$\begin{array}{l} \text{Give orthogonal } Q_0 \text{ such that } H = Q_0^T A Q_0 \text{ is upper Hessenberg.} \\ \text{For } k = 1, 2, \cdots, \\ \qquad H_i - h_{nn} I = Q_i R_i, \quad \text{(QR factorization)} \\ \qquad H_{i+1} := R_i Q_i + h_{nn} I, \\ \text{End} \end{array} \tag{5.4.18}$$

Quadratic convergence

If the $(n, n-1)$ entry converges to zero and let

$$H = \begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \varepsilon & h_{nn} \end{bmatrix},$$

then one step of the single shift QR algorithm leads:

$$QR = H - h_{nn}I, \;\; \bar{H} = RQ + h_{nn}I.$$

After $n-2$ steps in the reduction of $H - h_{nn}I$ to upper triangular we have

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & a & b \\ & & & \varepsilon & 0 \end{bmatrix}.$$

And we have $(n, n-1)$ entry in $\bar{H}$ is given by

$$\bar{h}_{n,n-1} = \frac{\varepsilon^2 b}{\varepsilon^2 + a^2}\ .$$

If $\varepsilon \ll a$, then it is clear that $(n, n-1)$ entry has order ε^2.

5.4.3 Double Shift QR iteration

If at some stage the eigenvalues a_1 and a_2 of $\begin{bmatrix} h_{mm} & h_{mn} \\ h_{nm} & h_{nn} \end{bmatrix}$ $(m = n-1)$ are complex, for then h_{nn} would tend to be a poor approximate eigenvalue. A way around this difficulty is to perform two single shift QR steps in succession, using a_1 and a_2 as shifts:

$$\begin{aligned} H - a_1 I &= Q_1 R_1, \\ H_1 &= R_1 Q_1 + a_1 I, \\ H_1 - a_2 I &= Q_2 R_2, \\ H_2 &= R_2 Q_2 + a_2 I. \end{aligned} \tag{5.4.19}$$

We then have

$$\begin{aligned} (Q_1 Q_2)(R_2 R_1) &= Q_1(H_1 - a_2 I)R_1 = Q_1(R_1 Q_1 + a_1 I - a_2 I)R_1 \\ &= (Q_1 R_1)(Q_1 R_1) + a_1(Q_1 R_1) - a_2(Q_1 R_1) \\ &= (H - a_1 I)(H - a_1 I) + a_1(H - a_1 I) - a_2(H - a_1 I) \\ &= (H - a_1 I)(H - a_2 I) = M, \end{aligned} \tag{5.4.20}$$

where

$$M = (H - a_1 I)(H - a_2 I). \tag{5.4.21}$$

Note that M is a real matrix, since

$$M = H^2 - sH + tI,$$

where $s = a_1 + a_2 = h_{mm} + h_{nn} \in \mathbb{R}$ and $t = a_1 a_2 = h_{mm}h_{nn} - h_{mn}h_{nm} \in \mathbb{R}$. Thus, (5.4.20) is the QR factorization of a real matrix, and we may choose Q_1 and Q_2 so that $Z = Q_1 Q_2$ is real orthogonal. It follows that

$$H_2 = Q_2^* H_1 Q_2 = Q_2^*(Q_1^* H Q_1)Q_2 = (Q_1 Q_2)^* H (Q_1 Q_2) = Z^T H Z$$

is real. A real H_2 could be guaranteed if we

(a) explicitly form the real matrix $M = H^2 - sH + tI$;

(b) compute the real QR decomposition $M = ZR$ and

(c) set $H_2 = Z^T H Z$.

But since (a) requires $O(n^3)$ flops, this is not a practical course. In light of the Implicit Q theorem, however, it is possible to effect the transition from H to H_2 in $O(n^2)$ flops if we

(a′) compute Me_1, the first column of M;

(b′) determine Householder Matrix P_0 such that

$$P_0(Me_1) = \alpha e_1, \quad (\alpha \neq 0);$$

(c′) compute Householder matrices $P_1, \cdots, P_{n-2}$ such that if $Z_1 = P_0 P_1 \cdots P_{n-2}$ the $Z_1^T H Z_1$ is upper Hessenberg and the first column of Z and Z_1 are the same. If $Z^T H Z$ and $Z_1^T H Z_1$ are both unreduced upper Hessenberg, then they are essentially equal.

Since $Me_1 = (x, y, z, 0, \cdots, 0)^T$, where $x = h_{11}^2 + h_{12}h_{21} - sh_{11} + t$, $y = h_{21}(h_{11} + h_{22} - s)$, $z = h_{21}h_{32}$. So, a similarity transformation with P_0 only changes rows and columns 1, 2 and 3. Since $P_0^T H P_0$ has the form

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix},$$

it follows that

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix} \xrightarrow{P_1} \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix} \xrightarrow{P_2} \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \end{bmatrix}$$

$$\xrightarrow{P_3} \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \end{bmatrix} \xrightarrow{P_4} \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix}.$$

$P_k = diag(I_k, \bar{P}_k, I_{n-k-3}), \bar{P}_k$ is 3×3-Householder matrix. The applicability of Theorem 5.4.6 (Implicit Q-theorem) follows from that $P_k e_1 = e_1$, for $k = 1, \cdots, n-2$, and that P_0 and Z have the same first column. Hence $Z_1 e_1 = Z e_1$.

Algorithm 5.4.3 (Francis QR step) *Given $H \in \mathbb{R}^{n\times n}$ unreduced whose trailing 2×2 principal submatrix has eigenvalues a_1 and a_2, the following algorithm overwrites H with $Z^T H Z$, where $Z = P_1 \cdots P_{n-2}$ is a product of Householder matrices and $Z^T(H - a_1 I)(H - a_2 I)$ is upper triangular.*

Set

$m := n - 1;$
$s := h_{mm} + h_{nn};$
$t := h_{mm}h_{nn} - h_{mn}h_{nm};$
$x := h_n^2 + h_{12}h_{21} - sh_{11} + t;$
$y := h_{21}(h_{11} + h_{22} - s);$
$z := h_{21}h_{32};$

For $k = 0, \cdots, n-2,$

If $k < n-2$, then

Determine a Householder matrix $\bar{P}_k \in \mathbb{R}^{3\times 3}$ such that

$$\bar{P}_k \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ 0 \end{bmatrix};$$

Set

$$H := P_k H P_k^T, \ P_k = diag\left(I_k, \bar{P}_k, I_{n-k-3}\right);$$

else determine a Householder matrix $\bar{P}_{n-2} \in \mathbb{R}^{2\times 2}$ such that

$$\bar{P}_{n-2} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix};$$

Set

$$H := P_{n-2} H P_{n-2}^T, \ P_{n-2} = diag\left(I_{n-2}, \bar{P}_{n-2}\right);$$

End if

$x := h_{k+2,k+1};$
$y := h_{k+3,k+1};$
If $k < n-3$, then $z := h_{k+4,k+1};$

End for;

This algorithm requires $6n^2$ flops. If Z is accumulated into a given orthogonal matrix, an additional $6n^2$ flops are necessary.

Algorithm 5.4.4 (QR Algorithm) *Given $A \in \mathbb{R}^{n\times n}$ and a tolerance ε, this algorithm computes the real schur decomposition $Q^T A Q = T$. A is overwritten with the Hessenberg decomposition.*

Using Algorithm 5.4.2 to compute the Hessenberg decomposition

$$Q^T A Q = H,$$

where $Q = P_1 \cdots P_{n-2}$ and H is Hessenberg;

Repeat: Set to zero all subdiagonal elements that satisfy

$$|h_{i,i-1}| \leq \varepsilon \left(|h_{ii}| + |h_{i-1,i-1}| \right);$$

Find the largest non-negative q and the smallest non-negative p such that

$$H = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ 0 & H_{22} & H_{23} \\ 0 & 0 & H_{33} \end{bmatrix} \begin{matrix} p \\ n-p-q \\ q \end{matrix},$$

$$\begin{matrix} p & n-p-q & q \end{matrix}$$

where H_{33} is upper quasi-triangular and H_{22} is unreduced (Note: either p or q may be zero).

If $q = n$, then upper triangularize all 2×2 diagonal blocks in H that have real eigenvalues, accumulate the orthogonal transformations if necessary, and quit.

Apply a Francis QR-step to H_{22}:

$$H_{22} := Z^T H_{22} Z;$$

If Q and T are desired, then $Q := Q \ diag(I_p, Z, I_q)$;

Set $H_{12} := H_{12} Z$ and $H_{23} := Z^T H_{23}$;

Go To Repeat.

This algorithm requires $15n^3$ flops, if Q and T are computed. If only the eigenvalues are desired, then $8n^3$ flops are necessary.

5.4.4 Ordering Eigenvalues in the Real Schur From

If $Q^T A Q = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$ with $T_{11} \in \mathbb{R}^{p \times p}$ and $\sigma(T_{11}) \cup \sigma(T_{22}) = \phi$, then the first p columns of Q span the unique invariant subspace associated with $\sigma(T_{11})$. Unfortunately, the Francis iteration leads $Q_R^T A Q_F = T_F$ in which the eigenvalues appear somewhat randomly along the diagonal of T_F. We need a method for computing an orthogonal matrix Q_D such that $Q_D^T T_F Q_D$ is upper quasitriangular with appropriate eigenvalues ordering.

Let $A \in \mathbb{R}^{2 \times 2}$, suppose

$$Q_F^T A Q_F = T_F = \begin{bmatrix} \lambda_1 & t_{12} \\ 0 & \lambda_2 \end{bmatrix}, \quad \lambda_1 \neq \lambda_2.$$

Note that $T_F x = \lambda_2 x$, where $x = \begin{bmatrix} t_{12} \\ \lambda_2 - \lambda_1 \end{bmatrix}$. Let Q_D be a given rotation such that $Q_D^T x = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$. If $Q = Q_F Q_D$, then

$$\left(Q^T A Q \right) e_1 = Q_D^T T_F \left(Q_D e_1 \right) = \lambda_2 Q_D^T \left(Q_D e_1 \right) = \lambda_2 e_1$$

and so

$$Q^T A Q = \begin{bmatrix} \lambda_2 & \pm t_{12} \\ 0 & \lambda_1 \end{bmatrix}.$$

Using this technique, we can move any subset of $\sigma(A)$ to the top of $T's$ diagonal. See Algorithm 7, 6–1 pp.241 (Golub & Van Loan: Matrix Computations). The swapping gets a little more complicated when T has 2×2 blocks. See Ruhe (1970) and Stewart (1976).

Block Diagonalization

Let

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1q} \\ 0 & T_{22} & \cdots & T_{2q} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & T_{qq} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \\ \}n_q \end{matrix} \tag{5.4.22}$$

be a partitioning of some real Schur form $Q^T AQ = T \in \mathbb{R}^{n\times n}$ such that $\sigma(T_{11}), \cdots, \sigma(T_{qq})$ are disjoint. There exists a matrix Y such that $Y^{-1}TY = \text{diag}(T_{11}, \cdots, T_{qq})$. A practical procedure for determining Y is now given together with an analysis of Y's sensitivity as a function of the above partitioning.

Partition $I_n = [E_1, \cdots, E_q]$ conformally with T and define $Y_{ij} \in \mathbb{R}^{n\times n}$ as follows:

$$Y_{ij} = I_n + E_i Z_{ij} E_j^T, \quad i < j,\ Z_{ij} \in \mathbb{R}^{n_i \times n_j}.$$

It follows that if $Y_{ij}^{-1} T Y_{ij} = \bar{T} = \left(\bar{T}_{ij}\right)$ then T and $\bar{T}$ are identical except that

$$\begin{aligned} \bar{T}_{ij} &= T_{ii}Z_{ij} - Z_{ij}T_{jj} + T_{ij}, \\ \bar{T}_{ik} &= T_{ik} - Z_{ij}T_{jk} \quad (k = j+1, \cdots, q), \\ \bar{T}_{kj} &= T_{ki}Z_{ij} + T_{kj} \quad (k = 1, \cdots, i-1). \end{aligned}$$

This $\bar{T}_{ij}$ can be zeroed provided we have an algorithm for solving the Sylvester equation

$$FZ - ZG = C, \tag{5.4.23}$$

where $F \in \mathbb{R}^{p\times p}$, $G \in \mathbb{R}^{r\times r}$ are given upper quasi-triangular and $C \in \mathbb{R}^{p\times r}$.

Bartels and Stevart (1972): Let $C = [c_1, \cdots, c_r]$ and $Z = [z_1, \cdots, z_r]$ be column partitionings. If $g_{k+1,k} = 0$, then by comparing columns in (5.4.23) we find

$$Fz_k - \sum_{i=1}^{k} g_{ik} z_i = c_k.$$

Thus, once we know $z_1, \cdots, z_{k-1}$ then we can solve the quasi-triangular system

$$(F - g_{kk})z_k = c_k + \sum_{i=1}^{k-1} g_{ik} z_i \text{ for } z_k.$$

If $g_{k+1,k} \neq 0$, then z_k and z_{k+1} can be simultaneously found by solving the $2p \times 2p$ system

$$\begin{bmatrix} F - g_{kk}I & -g_{mk}I \\ -g_{km}I & F - g_{mm}I \end{bmatrix} \begin{bmatrix} z_k \\ z_m \end{bmatrix} = \begin{bmatrix} c_k \\ c_m \end{bmatrix} + \sum_{i=1}^{k-1} \begin{bmatrix} g_{ik} z_i \\ g_{im} z_i \end{bmatrix} \quad (m = k+1). \tag{5.4.24}$$

By reordering the equations according to permutation $(1, p+1, p+2, \cdots, p, 2p)$, a banded system is obtained that can be solved in $O(p^2)$ flops. The detail may be found in Bartel and Stewart (1972) and see algorithm 7.6–2, 6–3 pp.243 (Golub & Van Loan Matrix Computation).

Connection with variant inverse iteration

Now let $A \in \mathbb{C}^{n\times n}$. The QR algorithm with respect to the sequence $\{k_i\}_{i=1}^{\infty}$ of shift:

$$\begin{aligned} A_1 &= A, \\ (A_i - k_i I) &= Q_i R_i, \\ A_{i+1} &= R_i Q_i + k_i I, \quad P_i = Q_1 Q_2 \cdots Q_i. \end{aligned}$$

Theorem 5.4.8 *Let p_s denote the last column of P_s. The sequence $\{p_s\}_{s=1}^{\infty}$ is then created by the variant inverse iteration:*

$$\begin{aligned} &p_0 = e_n, \quad k_1 = p_p^T A p_0, \\ &\textit{for } s = 0, 1, 2, \cdots \\ &\quad \tilde{p}_{s+1} = (A^* - k_{s+1} I)^{-1} p_s, \quad r_{s+1} = \left(\tilde{p}_{s+1}^* \tilde{p}_{s+1}\right)^{-1/2}, \\ &\quad p_{s+1} = r_{s+1} \tilde{p}_{s+1}, \quad k_{s+2} = p_{s+1}^* A p_{s+1}. \end{aligned}$$

Proof: $AP_s = P_s A_{s+1}$ implies

$$\begin{aligned} P_{s+1} &= P_s Q_{s+1} R_{s+1} R_{s+1}^{-1} = P_s (A_{s+1} - k_{s+1} I) R_{s+1}^{-1} \\ &= (A - K_{s+1} I) P_s R_{s+1}^{-1} \end{aligned}$$

and therefore

$$P_{s+1} = (A^* - \bar{k}_{s+1} I)^{-1} P_s R_{s+1}^* \text{ (since } P_s^{-*} = P_s).$$

If we denote by r the last diagonal element of R_{s+1}, then $p_{s+1} = (A^* - \bar{k}_{s+1} I)^{-1} p_s r$. From $(A_{s+1} - k_{s+1} I)^{-*} R_{s+1}^* = Q_{s+1}$ follows that

$$R_{s+1} P_s^* (A - k_{s+1} I)^{-1} (A^* - \bar{k}_{s+1} I)^{-1} P_s R_{s+1}^* = I$$

and then $r = r_{s+1}$. ■

Deflation "Remove" a computed eigenvalue and eigenvector from a matrix.

(a) Deflation from Hotelling: A is symmetric and real. Let λ_1 and x_1 be the computed eigenvalue and eigenvector respectively, and $x_1^T x_1 = 1$. Then

$$B = A - \lambda_1 x_1 x_1^*$$

has the following relation

$$B x_j = A x_j - \lambda_1 x_1 x_1^T x_j = \begin{cases} \lambda_j x_j, & j \neq 1, \\ 0 \cdot x_j, & j = 1, \end{cases}$$

where $A x_j = \lambda_j x_j \quad j = 1, \cdots, n$. B has the eigenvalues $\{0, \lambda_2, \cdots, \lambda_n\}$.

(b) Deflation from Wielandt: Let A be arbitrary. We know the fact, that a left eigenvector y to μ and a right eigenvector x to λ for $\lambda \neq \mu$ are orthogonal:

$$0 = (y^T A)x - y^T(Ax) = \mu y^T x - \lambda y^T x = (\mu - \lambda)y^T x.$$

Let λ_1 and x_1 be the given eigenvalue and the eigenvector respectively. Let $u \neq 0$ be a vector with $u^T x_1 \neq 0$. Then

$$B = A - x_1 u^T.$$

From

$$Bx_1 = \lambda_1 x_1 - (u^T x_1)x_1 = (\lambda_1 - u^T x_1)x_1$$

follows that the eigenvalue λ_1 is transformed to $\lambda_1 - u^T x_1$. If $\lambda \neq \lambda_1$ an eigenvalue, then follows from $y^T A = \lambda y^T$ $(y \neq 0)$ and $y^T B = y^T A - (y^T x_1)u^T = \lambda y^T$ that λ is also an eigenvalue of B. But the right eigenvectors are changed.

(c) Deflation with similarity transformation

A is arbitrary. Let x_1, λ_1 be given with $Ax_1 = \lambda_1 x_1$. Find a matrix H such that $Hx_1 = ke_1 (k \neq 0)$. Then holds

$$HAH^{-1}Hx_1 = \lambda_1 Hx_1 \quad \text{and} \quad HAH^{-1}e_1 = \lambda_1 e_1.$$

That is HAH^{-1} has the form

$$HAH^{-1} = \left(\begin{array}{c|c} \lambda_1 & b^T \\ \hline 0 & B \end{array}\right).$$

B has the eigenvalues $\sigma(A) \setminus \{\lambda_1\}$.

5.5 LR, LRC and QR algorithms for positive definite matrices

(a). LR-algorithm: Given matrix A. Consider

$$\begin{array}{l} A_1 := A, \\ \text{for } i = 1, 2, 3, \cdots \\ \quad A_i = L_i R_i, \quad (LR\text{factorization of } A_i) \\ \quad A_{i+1} := R_i L_i. \end{array} \tag{5.5.1}$$

From (5.4.5)–(5.4.7) we have

$$P_i := L_1 \cdots L_i, \; S_i := R_i \cdots R_1,$$

$$A_{i+1} := P_i^{-1} A P_i = S_i A S_i^{-1}, \tag{5.5.2}$$

$$A^i = P_i S_i. \tag{5.5.3}$$

There exists the convergence theorem as Theorem 5.4.2.
Advantage: less cost of computation at each step.
Disadvantage: LR factorization does not always exist.

(b). *LRC*-algorithm:

Let A be symmetric positive definite. Then the LR factorization exists. So we have the following iterative algorithm

$$\begin{array}{l} A_1 := A, \\ \text{for } i = 1, 2, 3, \cdots \\ \quad A_i = L_i L_i^T, \quad \text{(Cholesky factorization of } A_i \text{)} \\ \quad A_{i+1} := L_i^T L_i. \end{array} \tag{5.5.4}$$

Similar to (5.4.5)–(5.4.7) we also have

$$P_i := L_1 L_2 \cdots L_i,$$
$$A_{k+1} = P_k^{-1} A P_k = P_k^T A P_k^{-T}, \tag{5.5.5}$$
$$A^k = P_k P_k^T. \tag{5.5.6}$$

Because all P_i are positive definite, the LRC algorithm is always performable.

Theorem 5.5.1 *Let A be symmetric positive definite with eigenvalues $\lambda_1, \cdots, \lambda_n$. Then the LRC algorithm converges: The sequence A_k converges to a diagonal matrix Λ with the eigenvalues of A on the diagonal. If $\Lambda = diag(\lambda_i)$, where $\lambda_1 > \lambda_2 > \cdots \lambda_n > 0$, $A = U\Lambda U^T$ and U^T has a LR factorization, then A_k converges to Λ.*

Proof: Let $L_k = \left(\ell_{ij}^k\right)$ and $s_m^k = \sum_{i=1}^m a_{ii}^k, 1 \le m \le n$. Since all A_k are positive definite and $a_{ii}^k > 0$, we have

$$0 \le s_m^k \le \sum_{i=1}^n a_{ii}^k = \text{trace of } A_k = \text{trace of } A.$$

Thus s_m^k are bounded. From $A_k = L_k L_k^T$ follows $a_{ii}^k = \sum_{p=1}^i |\ell_{ip}^k|^2$. From $A_{k+1} = L_k^T L_k$ follows $a_{ii}^{k+1} = \sum_{p=i}^i |\ell_{pi}^k|^2$. Hence $s_m^k = \sum_{i=1}^m \sum_{p=1}^i |\ell_{ip}^k|^2$ and $s_m^{k+1} = \sum_{i=1}^m \sum_{p=i}^n |\ell_{pi}^k|^2$.

The skizze shows clearly that $s_m^{k+1} \ge s_m^k$. So s_m^k converges, and then $a_{ii}^k = s_i^k - s_{i-1}^k$ and $s_m^{k+1} - s_m^k = \sum_{p=1}^m \sum_{j=m+1}^n \left(\ell_{jp}^k\right)^2 \to 0$. This shows that $\ell_{pj}^k \to 0,\ p \ne j$ and since $a_{ii}^k = \left(\ell_{ii}^k\right)^2 + \sum_{p=1}^{i-1} \left(\ell_{ip}^k\right)^2$ and $a_{ii}^k > 0$, so ℓ_{ii}^k converges. So L_i converges to a diagonal matrix. Here $A_i = L_i L_i^T$.

Second part: From $A = U\Lambda U^T$, $U^T = LR$ follows

$$\begin{aligned} A^s &= U\Lambda^s U^T = R^T L^T \Lambda^s L R \qquad (s = 2t) \\ &= R^T \Lambda^t \left(\Lambda^{-t} L^T \Lambda^t\right) \left(\Lambda^t L \Lambda^{-t}\right) \Lambda^t R. \end{aligned}$$

Since $\Lambda^t L \Lambda^{-t} = I + E_t$ with $E_t \to 0$ and by continuity of LL^T-factorization we have

$$\left(\Lambda^{-t} L^T \Lambda^t\right) \left(\Lambda^t L^T \Lambda^{-t}\right) = (I + E_t)^T (I + E_t) = \tilde{L}_s \tilde{L}_s^T, \quad \tilde{L}_s \to I$$

and

$$A^s = R^T \Lambda^t \tilde{L}_s \cdot \tilde{L}_s \Lambda^t R = P_s P_s^T.$$

We now have two different LL^T-decomposition of A^s. There is a unitary diagonal matrix D_s with

$$P_s D_s = R^T \Lambda^T \tilde{L}_s$$

and hence

$$\begin{aligned} D_s^{-1}A_{s+1}D_s &= D_s^{-1}P_s^{-1}AP_sD_s \\ &= \tilde{L}_s^{-1}\Lambda^{-t}\left(R^{-T}AR^T\right)\Lambda^t\tilde{L}_s \\ &= \tilde{L}_s^{-1}\Lambda^{-t}\left(L^T\Lambda L^{-T}\right)\Lambda^t\tilde{L}_s. \end{aligned}$$

Since $A = U\Lambda U^{-1} = R^TL^T\Lambda L^{-T}R^{-T}$ and $L^T\Lambda L^{-T}$ is a upper triangular with diagonal Λ, it holds $\Lambda^{-t}L^T\Lambda L^{-T}\Lambda^t \to \Lambda$ and because of $\tilde{L}_s \to I$, it holds $D_s^{-1}A_{s+1}D_s \to \Lambda$, also $A_{s+1} \to \Lambda$. ■

Remark 5.5.1 (i) *One can also develop shift-strategy and deflation technique for LR and LRC algorithm as in QR algorithm.*

(ii) *If A is a (k,k)-band matrix, then L_1 is a $(k,0)$-band matrix and therefore $A_2 = L_1^TL_1$ is also a (k,k)-band matrix. The band structure is preserved.*

(c). **QR-algorithm for positive definite matrices**

We apply QR-algorithm (5.4.3)–(5.4.4) to symmetric matrices. From

$$A_{i+1} = Q_i^*A_iQ_i$$

follows that A_i are symmetric.

Theorem 5.5.2 *The QR algorithm converges for positive definite matrices.*

The proof follows immediately from the following Theorem 5.5.3.

We consider now the iteration of QR algorithm

$$A_{i+1} = Q_i^*A_iQ_i$$

and the iterations of LRC algorithm

$$\tilde{A}_i := L_iL_i^T, \quad \tilde{A}_{i+1} = L_i^TL_i.$$

Theorem 5.5.3 *The $(i+1)$-th iteration A_{i+1} of QR algorithm for positive definite A corresponds to the $(2i+1)$-th iteration $\tilde{A}_{2i+1}$ of LRC algorithm for $i = 0, 1, 2, \cdots$.*

Proof: From (5.4.5)–(5.4.7) we have

$$P_i := Q_i\cdots Q_i, \;\; S_i := R_i\cdots R_1 \tag{5.5.7}$$

and

$$A^i = P_iS_i, \quad A_{i+1} = S_iAS_i^{-1}. \tag{5.5.8}$$

Similarly, from (5.5.2) and (5.5.3) with $\tilde{P}_i = L_1\cdots L_i$, we have

$$A^i = \tilde{P}_i\tilde{P}_i^T, \quad \tilde{A}_{i+1} = \tilde{P}_i^TA\tilde{P}_i^{-T}. \tag{5.5.9}$$

From(5.5.8) follows

$$A^{2i} = (A^i)^*A^i = S_i^*P_i^*P_iS_i = S_i^*S_i.$$

On the other hand from (5.5.9) with $i \leftarrow 2i$ follows

$$A^{2i} = \tilde{P}_{2i}\tilde{P}_{2i}^T.$$

From the uniqueness of LRC factorization of positive diagonal follows $S_i = \tilde{P}_{2i}^T$ and hence according to (5.5.8) (5.5.9) it holds

$$A_{i+1} = S_i A S_i^{-1} = \tilde{P}_{2i}^T A \tilde{P}_{2i}^{-T} = \tilde{A}_{2i+1}.$$

■

The proof of Theorem 5.5.2 is now from Theorem 5.5.1 and Theorem 5.5.3 evident.

Remark 5.5.2 *For positive definite matrices two steps of LL^T algorithm are as many as one step of QR algorithm. This shows that QR algorithm is much more favorable.*

5.6 *qd*-algorithm (Quotient Difference)

We indicated in Remark 5.5.1(ii), the band structure is preserved by LR algorithm. Let $A = A_1$ be a (k,m)-band matrix. Then all $L_i, (k,0)-$, all $R_i, (0,m)-$ and all $A_i, (k,m)$-band matrices, respectively. Especially tridiagonal form is preserved. A transformation of LR-algorithm for tridiagonal matrices derives to qd-algorithm. A tridiagonal matrix

$$\tilde{A} = \begin{pmatrix} \tilde{\alpha}_1 & \tilde{\beta}_2 & & & 0 \\ \tilde{\gamma}_2 & \tilde{\alpha}_2 & \tilde{\beta}_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \tilde{\beta}_n \\ 0 & & & \tilde{\gamma}_n & \tilde{\alpha}_n \end{pmatrix} \tag{5.6.1}$$

for $\tilde{\beta}_i \neq 0$ $(i = 2, \cdots, n)$ can be transformed with $D = \mathrm{diag}(1, \tilde{\beta}_2, \tilde{\beta}_2\tilde{\beta}_3, \cdots, \tilde{\beta}_2 \cdots \tilde{\beta}_n)$ to the form $D\tilde{A}D^{-1} = A$, where

$$A = \begin{pmatrix} \alpha_1 & 1 & & & 0 \\ \gamma_1 & \alpha_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & \gamma_n & \alpha_n \end{pmatrix} \tag{5.6.2}$$

with $\gamma_i = \tilde{\beta}_i\tilde{\gamma}_i$ and $\alpha_i = \tilde{\alpha}_i$. Hence without loss of generality we can study the form (5.6.2) for tridiagonal matrices. We now apply LR-algorithm to (5.6.2):

$$A_s = \begin{pmatrix} \alpha_1^s & 1 & & 0 \\ \gamma_2^s & \alpha_2^s & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & \gamma_n^s & \alpha_n^s \end{pmatrix}, \; L_s = \begin{pmatrix} 1 & & & 0 \\ e_2^s & \ddots & & \\ & \ddots & \ddots & \\ 0 & & e_n^s & 1 \end{pmatrix},$$

$$R_s = \begin{pmatrix} q_1^s & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & q_n^s \end{pmatrix}. \tag{5.6.3}$$

The LR factorization $A_s = L_s R_s$ can be obtained by element comparison:

$$\begin{array}{rll} (1,1): & \alpha_1^s = q_1^s, & \\ (i,i-1): & \gamma_i^s = e_i^s q_{i-1}^s, & i = 2, \cdots, n, \\ (i,i): & \alpha_i^s = e_i^s + q_i^s, & i = 2, \cdots, n, \\ (i,i+1): & 1 = 1 \cdot 1, & i = 1, \cdots, n-1. \end{array} \tag{5.6.4}$$

We can determine e_i^s, q_i^s from above equations for a given A_s in the sequence $q_1^s, e_2^s, q_2^s, e_3^s, q_3^s, \ldots, q_n^s$ and compute $A_{s+1} = R_s L_s$ by

$$\begin{cases} \alpha_i^{s+1} = q_i^s + e_{i+1}^s, & i = 1, \cdots, n-1 \\ \alpha_n^{s+1} = q_n^s, & \\ \gamma_i^{s+1} = q_i^s e_i^s, & i = 2, \cdots, n. \end{cases} \tag{5.6.5}$$

We write $s+1$ instead of s in (5.6.4), then we can eliminate A_{s+1} and obtain

$$\begin{cases} (\alpha_i^{s+1} =) e_i^{s+1} + q_i^{s+1} = q_i^s + e_{i+1}^s, & i = 1, \cdots, n \\ (\gamma_i^{s+1} =) e_i^{s+1} q_{i-1}^{s+1} = q_i^s e_i^s, & i = 2, \cdots, n \end{cases}. \tag{5.6.6}$$

For the convenience of notation we suppose

$$e_1^s = 0, \quad e_{n+1}^s = 0, \quad s = 1, 2, \cdots. \tag{5.6.7}$$

The equations (5.6.6) can be represented by the qd-scheme and the Rhomben rules:

qd-Scheme

$$\begin{array}{r|llllllll|l} (e_1^s =)\, 0 & q_1^s & e_2^s & & & & & & & \\ \left(e_1^{s+1} =\right) 0 & q_1^{s+1} & e_2^{s+1} & q_2^s & & & & & & \\ & & & q_2^{s+1} & \ddots & q_{n-1}^s & e_n^s & & & \\ & & & & \ddots & q_{n-1}^{s+1} & e_n^{s+1} & q_n^s & & 0\ \left(= e_{n+1}^s\right) \\ & & & & & & & q_n^{s+1} & & 0\ \left(= e_{n+1}^{s+1}\right) \end{array}$$

The first equations in (5.6.6) can be formulated as sum rule:

$$\begin{array}{ccccc} & & q_i^s & & \\ & & & \ddots & \\ e_i^{s+1} & & & & e_{i+1}^s \\ & \ddots & & & \\ & & q_i^{s+1} & & \end{array}$$

The sum of elements of upper rows is equal to the sum of elements of lower rows. Thus,

$$q_i^{s+1} = q_i^s + e_{i+1}^s - e_i^{s+1}. \tag{5.6.8}$$

The second equations in (5.6.6) can be formulated as product rule:

$$\begin{array}{ccccc} & & e_i^s & & \\ & & & \ddots & \\ q_{i-1}^{s+1} & & & & q_i^s \\ & \ddots & & & \\ & & e_i^{s+1} & & \end{array}$$

The product of elements of upper rows is equal to the product of elements of lower rows. Thus,

$$e_i^{s+1} = \frac{e_i^s q_i^s}{q_{i-1}^{s+1}}. \tag{5.6.9}$$

With these rules a new qd-rows can be determined by sum and product rules from left to right. Start according to (5.6.4) with $s = 1$. The formulas (5.6.8)(5.6.9) interpret the name quotient-difference algorithm.

5.6.1 The qd-algorithm for positive definite matrix

If $\tilde{A}$ in (5.6.1) is positive definite, then $\det \tilde{A} > 0$, and it also holds for A because $\det A = \det D \det \tilde{A} \det D^{-1} = \det \tilde{A} > 0$. This also holds for principal determinants $h_1, \cdots, h_n$ of $\tilde{A}$. They are positive and equal to principal determinants of A, respectively. In general we have

Lemma 5.6.1 *If a matrix B is diagonal similar to a positive definite matrix C, then all principal determinants of B are positive.*

Lemma 5.6.2 *A matrix in the form (5.6.2) is diagonal similar to a symmetric tridiagonal matrix, if and only if, $\gamma_i > 0$, for $i = 2, \cdots, n$. Especially this matrix is irreducible.*

Proof: If $\gamma_i > 0$, then $D^{-1}AD$ is symmetric, where

$$D = \mathrm{diag}(1, t_2, t_2 t_3, \cdots, t_2 \cdots t_n), \quad t_i := \sqrt{\gamma_i}\ .$$

Reversely, if D is a diagonal matrix, $D = diag(d_i)$ and $\tilde{A} = D^{-1}AD$ symmetric, then $\tilde{a}_{i,i+1} = d_{i+1}/d_i = \tilde{a}_{i+1,i} = \gamma_i(d_i/d_{i+1})$ and $d_{i+1}/d_i \neq 0$. So $\gamma_i = (\tilde{a}_{i,i+1})^2 > 0$. ■

Theorem 5.6.1 *The qd-algorithm converges for irreducible, symmetric positive definite tridiagonal matrices. i.e. If $\tilde{A}$ is irreducible and positive definite, then it holds the quantities computed from (5.6.2) (5.6.4) (5.6.8)(5.6.9):*

$$e_i^s > 0, \quad \lim_{s\to\infty} e_i^s = 0, \quad i = 2, \cdots, n, \tag{5.6.10}$$

$$q_i^s > 0, \quad \lim_{s\to\infty} q_i^s = 0, \quad i = 1, \cdots, n. \tag{5.6.11}$$

Hereby $\lambda_i, i = 1, \cdots, n$ are the eigenvalues of $\tilde{A}$ and satisfy

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0. \tag{5.6.12}$$

Proof: Let h_i^k be the i-th principal determinant of A_k. We first show that by induction on k:

$$e_i^k > 0,\ i = 2, \cdots, n,\ q_i^k > 0,\ h_i^k > 0, i = 1, \cdots, n.$$

For $A = A_1$, Lemma 5.6.1 shows that: $h_i^1 > 0, i = 1, \cdots, n$. In addition we have from $A_s = L_s R_s$ that

$$h_i^s = q_i^s \cdots q_i^s, \quad i = 1, \cdots, n. \tag{5.6.13}$$

Hence for $s = 1 : q_i^1 > 0, i = 1, \cdots, n$. From Lemma 5.6.2 follows $\gamma_i = \gamma_i^1 > 0, i = 2, \cdots, n$, so from (5.6.4) we get

$$e_i^1 = \gamma_i^1 \Big/ q_{i-1}^1 > 0.$$

We suppose the above assertion is true until $k-1$, then from (5.6.5) follows

$$\gamma_i^k = q_i^{k-1} e_i^{k-1} > 0,$$

so from Lemma 5.6.2 and Lemma 5.6.1 we have that A_k is diagonal similar to a symmetric matrix, which must be positive definite, because A_k is similar to $\tilde{A}$. Hence all $h_i^k > 0$. Therefore from (5.6.13) q_i^k and from (5.6.4) e_i^k are also positive.

We now show that

$$\lim_{k\to\infty} e_i^k = 0, \ \lim_{k\to\infty} q_i^k = q_i > 0.$$

From (5.6.6) for $i = n, q_n^{s+1} + e_n^{s+1} = q_n^s$ follows that q_n^s is monotone decreasing, so q_n^s converges and $e_n^{s+1} = q_n^s - q_n^{s+1}$ approaches to zero. Adding the following equations together

$$\begin{aligned}
q_n^{k+1} &= q_n^k - e_n^{k+1}, \\
q_{n-1}^{k+2} &= q_{n-1}^{k+1} + e_n^{k+1} - e_{n-1}^{k+2}, \\
&\vdots \\
q_{n-\nu}^{k+\nu+1} &= q_{n-\nu}^{k+\nu} + e_{n+1-\nu}^{k+\nu} - e_{n-\nu}^{k+\nu+1},
\end{aligned}$$

we get that

$$q_n^{k+1} + q_{n-1}^{k+2} + \cdots + q_{n-\nu}^{k+\nu+1} = q_n^k + q_{n-1}^{k+1} + \cdots + q_{n-\nu}^{k+\nu} - e_{n-\nu}^{k+\nu+1},$$

i.e.,

$$\rho_\nu^{k+1} = \rho_\nu^k - e_{n-\nu}^{k+\nu+1}.$$

The sequence ρ_ν^k is positive, monotone decreasing, so it converges, for $\nu = 1, \cdots, n-1$. Hence q_ν^k converges to a number $q_\nu \geq 0$, thus $\lim_{k\to\infty} e_\nu^k = 0$. So $\lim_{s\to\infty} L_s = I$ and hence

$$\lim_{s\to\infty} A_s = \lim_{s\to\infty} L_s R_s = \lim_{s\to\infty} R_s = \begin{bmatrix} q_1 & 1 & & 0 \\ & \ddots & \ddots & 1 \\ & & \ddots & \\ 0 & & & q_n \end{bmatrix}.$$

This shows that q_i are the eigenvalues of A and $\tilde{A}$. It is necessary to show that q_i are in decreasing order. Suppose $q_i/q_{i-1} > 1$ for one i, then also holds for all s, $q_i^s/q_{i-1}^s > 1$. This contradicts that

$$e_i^{s+1} = e_i^s q_i^s / q_{i-1}^s \text{ and } e_i^s \to 0.$$

On the other hand, $q_i = q_{i-1}$ is not possible, since a tridiagonal matrix with nonzero subdiagonal only possesses simple eigenvalues. ■

Remark 5.6.1 *It is remarkable that the qd-algorithm has the advanced applications in the numerical mathematics for the computation of roots of polynomials.*

Chapter 6

The Symmetric Eigenvalue problem

6.1 Properties, Decomposition, Perturbation Theory

$$\text{A Hermitian} \Longleftrightarrow A = A^* \Longleftrightarrow A = (a_{ik}), a_{ik} = \bar{a}_{ki}, i, k = 1, \cdots, n.$$
$$\text{A symmetric} \Longleftrightarrow A = \bar{A}, A = A^T \Longleftrightarrow a_{ik} = a_{ki}, a_{ik} = \bar{a}_{ik}, i, k = 1, \cdots, n.$$

Theorem 6.1.1 (Schur Decomposition for Hermitian matrices) *If $A \in \mathbf{C}^{n\times n}$ is Hermitian (real symmetric), then there exists a unitary (orthogonal) Q such that*

$$\begin{aligned} &Q^*AQ = \wedge \equiv diag(\lambda_1, \cdots, \lambda_n), \\ &Aq_i = \lambda_i q_i, \quad i = 1, \cdots, n, \quad Q = [q_1, \cdots, q_n]. \end{aligned} \tag{1.1}$$

Proof: Let $Q^*AQ = T$ be the Schur Decomposition of A. It follows that T must be a direct sum of 1×1 and 2×2 matrices, since T is Hermitian. But a 2×2 Hermitian matrix can not have complex eigenvalues. Consequently, T has no 2×2 block along its diagonal. ■

Classical techniques:

There are extremely effetive techniques based on the minimax principle, for investigating the eigenvalues of the sum of two symmetric matrices.

Let X be a symmetric matrix defined by

$$X = \left[\begin{array}{c|c} \alpha & a^T \\ \hline a & diag(\alpha_i) \end{array}\right] \quad (i = 1, \cdots, n).$$

We wish to relate the eigenvalues of X with the α_i.

Suppose that only s of the components of a are non-zero. If a_j is zero, then α_j is an eigenvalue of X. There exists a permutation P such that

$$Y = P^T X P = \left(\begin{array}{c|c|c} \alpha & b^T & 0 \\ \hline b & diag(\beta_i) & 0 \\ \hline 0 & 0 & diag(\gamma_i) \end{array}\right),$$

where no component of b is zero, $diag(\beta_i)$ is of order s, and $diag(\gamma_i)$ is of order $n-1-s$.

The eigenvalues of X are therefore γ_i together with those of the matrix Z defined by

$$Z = \left(\begin{array}{c|c} \alpha & b^T \\ \hline b & diag(\beta_i) \end{array} \right).$$

If $s = 0$, Z is the single element α and hence the eigenvalues of X are $diag(\alpha_i)$ and α. Otherwise examine the characteristic polynominal of Z:

$$(\alpha - \lambda) \prod_{i=1}^{s} (\beta_i - \lambda) - \sum_{j=1}^{s} b_j^2 \prod_{i \neq j} (\beta_i - \lambda) = 0. \tag{1.1.1}$$

Suppose that there are only t distinct values among the β_i. *W.l.o.g.* we may take them to be $\beta_1, \cdots, \beta_t$ with multiplicities $r_1, r_2, \cdots, r_t$ respectively, so that $r_1 + r_2 + \cdots + r_t = s$. Clearly the left-hand side of (1.1.1) has the factor

$$\prod_{i=1}^{t} (\beta_i - \lambda)^{r_i - 1},$$

so that β_i is an eigenvalue of Z of multiplicity $(r_i - 1)$.

Dividing (1.1.1) by $\Pi_{i=1}^{t} (\beta_i - \lambda)^{r_i}$ we see that the remaining eigenvalues of Z are the roots of

$$0 = (\alpha - \lambda) - \sum_{i=1}^{t} c_i^2 (\beta_i - \lambda)^{-1} \equiv \alpha - f(\lambda), \tag{1.1.2}$$

where c_i^2 is the sum of the r_i values b_j^2 associated with β_i and is therefore strictly positive. A graph of $f(\lambda)$ against λ is given as follows, where it is assumed that distinct β_i are in decreasing order.
It is immediately evident that the $t+1$ roots of $\alpha = f(\lambda)$ which we denote by $\delta_1, \delta_2, \cdots, \delta_{t+1}$ satisfy

$$\infty > \delta_1 > \beta_1; \ \beta_{i-1} > \delta_i > \beta_i \ (i = 2, 3, \cdots, t); \ \beta_t > \delta_{t+1} > -\infty \tag{1.1.3}$$

The n eigenvalues of X therefore fall into three sets:

(1) The eigenvalues $\gamma_1, \cdots, \gamma_{n-1-s}$ corresponding to the zero a_i. These are equal to $n - 1 - s$ of the α_i.

(2) $s - t$ eigenvalues consisting of $r_i - 1$ values equal to β_i $(i = 1, 2, \cdots, t)$. These are equal to a further $s - t$ of the α_i.

(3) $t + 1$ eigenvalues equal to δ_i satisfying (1.1.3). If $t = 0$ then $\delta_1 = \alpha$.

Let the eigenvalues of X be denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then it is an immediate consequence of our enumeration of the λ_i above that if the α_i are also arranged in nonincreasing order then

$$\lambda_1 \geq \alpha_1 \geq \lambda_2 \geq \alpha_2 \geq \cdots \geq \alpha_{n-1} \geq \lambda_n. \tag{1.1.4}$$

In other words the α_i separate the λ_i at least in the weak sense.

Consider now the eigenvalues of X' derived from X by replacing α and α'. The eigenvalues of X' will equal to those of X as far as sets (1) and (2) are concerned.

Let us denote those in (3) by $\delta_1', \delta_2', \cdots, \delta_{t+1}'$. Now for $\lambda > 0$, we have

$$\frac{df}{d\lambda} = 1 + \sum_{i=1}^{t} \frac{c_i^2}{(\beta_i - \lambda)^2} > 1, \tag{1.1.5}$$

and hence $\delta_i{}' - \delta_i$ lies between 0 and $\alpha' - \alpha$. We may write

$$\delta_i{}' - \delta_i = m_i(\alpha' - \alpha), \tag{1.1.6}$$

where $0 \leq m_i \leq 1$ and $\sum_{i=1}^{t+1} m_i = 1$. If $t = 0$ then $\delta_1{}' = \alpha'$ and $\delta_1 = \alpha$ and $\delta_1{}' - \delta_1 = \alpha' - \alpha$. Hence we may write in all cases

$$\delta_i{}' - \delta_i = m_i(\alpha' - \alpha),$$

where $0 \leq m_i \leq 1$ and $\sum_{i=1}^{t+1} m_i = 1$. Since the other eigenvalues of X and X' are equal, we have established a correspondence between n eigenvalues $\lambda_1, \cdots, \lambda_n$ and $\lambda_1{}', \cdots, \lambda_n{}'$ of X and X' respectively.

$$\begin{aligned} &\lambda_i{}' - \lambda_i = m_i(\alpha' - \alpha), \\ &0 \leq m_i \leq 1, \quad \sum_{i=1}^{n} m_i = 1, \end{aligned} \tag{1.1.7}$$

where $m_i = 0$ for the eigenvalues from sets (1) and (2).

Now let $C = A + B$, where A and B are symmetric and B is of rank 1. There exists an orthogonal matrix R such that $R^T B R = \left[\begin{array}{c|c} \rho & 0 \\ \hline 0 & \bigcirc \end{array}\right]$, $\rho \neq 0$. Let

$$R^T A R = \left[\begin{array}{c|c} \alpha & a^T \\ \hline a & A_{n-1} \end{array}\right].$$

Then there is an orthogonal matrix S of order n-1 such that

$$S^T A_{n-1} S = diag(\alpha_i),$$

and if we define Q by

$$Q = R \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & S \end{array}\right],$$

then Q is orthogonal and

$$Q^T(A + B)Q = \left[\begin{array}{c|c} \alpha & b^T \\ \hline b & diag(\alpha_i) \end{array}\right] + \left[\begin{array}{c|c} \rho & 0 \\ \hline 0 & \bigcirc \end{array}\right],$$

where $b = S^T a$, the eigenvalues of A and of $(A + B)$ are therefore those of

$$\left[\begin{array}{c|c} \alpha & b^T \\ \hline b & diag(\alpha_i) \end{array}\right] \text{ and } \left[\begin{array}{c|c} \alpha + \rho & b^T \\ \hline b & diag(\alpha_i) \end{array}\right],$$

and if we denote these eigenvalues by λ_i and $\lambda_i{}'$ in decreasing order, then they satisfy

$$\lambda_i^{'} - \lambda_i = m_i\rho, \tag{1.1.8}$$
$$0 \le m_i \le 1 \quad \sum_{i=1}^{n} m_i = 1.$$

Hence when B is added to A, all eigenvalues of the latter are shifted by an amount which lies between zero and the eigenvalue ρ of B. We summary the above discussion as the following theorem.

Theorem 6.1.2 *Supose $B = A + \tau\, c\, c^T$, where $A \in R^{n\times n}$ is symmetric, $c \in R^n$ has unit 2-norm and $\tau \in R$. If $\tau \ge 0$ then*

$$\lambda_i(B) \in [\lambda_i(A), \lambda_{i-1}(A)], \quad i = 2, 3, \ldots, n,$$

while if $\tau \le 0$ then

$$\lambda_i(B) \in [\lambda_{i+1}(A), \lambda_i(A)], \quad i = 1, 2, \ldots, n-1.$$

In either case

$$\lambda_i(B) = \lambda_i(A) + m_i\tau,$$

where $m_1 + m_2 + \cdots + m_n = 1$ and $m_i \ge 0$. ■

Let $\lambda_i(A)$ denote the ith largest eigenvalue of A. Then

$$\lambda_n(A) \le \lambda_{n-1}(A) \le \cdots \le \lambda_1(A). \tag{1.2}$$

Definition 6.1.1 *If $A = A^*$,$x \ne 0$, then*

$$R[x] = \frac{x^T Ax}{x^T x}$$

is called the Rayleigh-Quotient of x, sometimes denoted by $R[x, A]$.

Theorem 6.1.3 *If $A = A^*$, then it holds*

$$\lambda_n(A) \le R[x] \le \lambda_1(A). \tag{1.3}$$

Proof: From (1.1) we have

$$R[x] = \frac{x^*Ax}{x^*x} = \frac{x^*UAU^*x}{x^*UU^*x} = \frac{y^*\Lambda y}{y^*y} = \frac{\sum_{i=1}^n \lambda_i |y_i|^2}{\sum_{i=1}^n |y_i|^2}, (y = U^*x). \tag{1.4}$$

Thus $R[x]$ is a convex combination of λ_i, it follows (1.3). ■

Corollary 6.1.4

$$\lambda_1(A) = \max_{x\ne 0} R[x] \quad \textit{and} \quad \lambda_n(A) = \min_{x\ne 0} R[x].$$

Theorem 6.1.5 (Weyl) *If A is Hermitian with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and eigenvectors $u_1, \cdots, u_n$, then it holds*

$$\lambda_i = \max\{R[x] : x \neq 0, x \perp u_k, k = 1, \cdots, i-1\}, \tag{1.5}$$

for $i = 1, \cdots, n$.

Proof: It is clear for $i = 1$. Let $i > 1$. If $x \perp u_1, \cdots, u_{i-1}$ then $u_k^* x = 0$, for $k = 1, \cdots, i-1$. So $y = U^* x$ satisfies $y_k = 0, k = 1, \cdots, i-1$ (Here $U = [u_1, \cdots, u_n]$). It follows from (1.4) that

$$R[x] = \frac{\sum_{j=i}^n \lambda_j \|y_j\|^2}{\sum_{j=i}^n \|y_j\|^2} \leq \lambda_i.$$

For $x = u_i$, we have $R[x] = \lambda_i$, so (1.5) holds. ■

Theorem 6.1.6 (Courant-Fischer) *Under above assumptions we have*

$$\lambda_i = \min_{\{p_1, \cdots, p_n\} l.i.} \{max\{R[x] : x \neq 0, x \perp p_k, k = 1, \cdots, i-1\}\} \tag{1.6}$$

$$\lambda_i = \min_{dim S = n+1-i} \{max\{R[x] : x \in S \setminus \{0\}\}\}. \tag{1.7}$$

$$\lambda_i = \max_{dim S = i} \{min\{R[x] : x \in S \setminus \{0\}\}\}. \tag{1.8}$$

Proof:

(1.6)$\Longleftrightarrow$ (1.7) trivial.

(1.7) $\Rightarrow$ (1.8): Applying (1.7) to $-A$, we then have

$$\lambda_i(-A) = \min_{dim S = n+1-i} \{max\{-R[x] : x \in S \setminus \{0\}\}\}.$$

That is

$$-\lambda_{n+1-i}(A) = -\max_{dim S = n+1-i} \{min\{-R[x] : x \in S \setminus \{0\}\}\}.$$

(Use $\max(-a_i) = -\min(a_i)$, $\min(-a_i) = -\max(a_i)$). By substituting $i \to n+1-i$ follows (1.8).

Claim (1.6): Since $\lambda_1 = \max_{x \neq 0}(R[x])$, for $i = 1$ it is true.

Consider $i > 1$: Let $p_1, \cdots, p_{i-1} \neq 0$ be given. The linear system

$$\begin{cases} {p_k}^* x = 0, & k = 1, \cdots, i-1, \\ {u_k}^* x = 0, & k = i+1, \cdots, n \end{cases}$$

has a solution $x \neq 0$, because of $n-1$ homogenous equations with n variables. Let $U = [u_1, \cdots, u_n]$. Then

$$R[x] = \frac{x^* U \Lambda U^* x}{x^* U U^* x} = \frac{\sum_{j=1}^i \lambda_j |U^* x|_j^2}{\sum_{j=1}^i |U^* x|_j^2} \geq \lambda_i.$$

But $p_k^* x = 0$, $k = 1, \cdots, i-1$ so

$$\max\{R[x] : \ x \perp p_k, \ k = 1, \cdots, i-1\} \geq \lambda_i.$$

This implies

$$\lambda_i \leq \min_{\{p_k\}_{k=1}^{i-1}} \{max\{R[x] : \ x \perp p_k, k = 1, \cdots, i-1\}\}.$$

Now set $p_k = u_k, k = 1, \cdots, i-1$. By (1.5) we have the equality (1.6). ■

Theorem 6.1.7 (Separation theorem) *A is Hermitian with eigenvalues $\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$. Let*

$$A_{n-1} = \begin{pmatrix} a_{11} & \cdots & a_{1,n-1} \\ \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n-1} \end{pmatrix}$$

be the $n-1$ principal submatrix of A with eigenvalues $\lambda_{n-1}^{'} \leq \cdots \leq \lambda_1{}'$. Then it holds $\lambda_{s+1} \leq \lambda_s^{'} \leq \lambda_s$, for $s = 1, \cdots, n-1$.

Proof: Let $z = \begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathbf{C}^n$, where $x \in \mathbf{C}^{n-1}$. Then

$$\frac{x^* A_{n-1} x}{x^* x} = \frac{z^* A z}{z^* z}.$$

Applying (1.5) to A_{n-1} we have

$$\begin{aligned} \lambda_s^{'} &= \max\{\frac{x^* A_{n-1} x}{x^* x} : 0 \neq x \in \mathbf{C}^{n-1}, x \perp u_i^{'}, i = 1, \cdots, s-1\} \\ &= \max\{\frac{z^* A z}{z^* z} : 0 \neq z \in \mathbf{C}^n, z \perp \begin{bmatrix} u_i^{'} \\ 0 \end{bmatrix}, \ e_n^T z = 0, i = 1, \cdots, s-1\} \\ &\geq \min_{\{p_i\}_{i=1}^{s} \ l.i.} \max\{R[z] : z \perp p_i, i = 1, \cdots, s\} = \lambda_{s+1} \quad (By(1.6)). \end{aligned}$$

therefore $\lambda_{s+1} \leq \lambda_s^{'}$. Here $u_i^{'}$ is the eigenvector of A_{n-1}. Now set $A \to -A$ then

$$\lambda_{s+1}(-A) \leq \lambda_s^{'}(-A_{n-1}).$$

Thus

$$-\lambda_{n-s}(A) \leq -\lambda_{n-s}^{'}(A_{n-1}).$$

It follows

$$\lambda_{n-s}(A) \geq \lambda_{n-s}^{'}(A_{n-1}).$$

Hence we have $\lambda_{n-s} \geq \lambda_{n-s}^{'}$. By setting $s \to n-s$, we have $\lambda_s \geq \lambda_s^{'}$. ■

Theorem 6.1.8 (Separation theorem) *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A and $\lambda_1{}' \geq \cdots \geq \lambda_{n-1}'$ be the eigenvalues of B', where B' is obtained by scratching a row and the same column of A, then $\lambda_{s+1} \leq \lambda_s{}' \leq \lambda_s$. Further consequence are: If B consists of by scratching two rows and the coresponding columns of B', i.e., $A \to B' \to B$, then we have*

$$\lambda_{i+2} \leq \lambda_{i+1}^{'} \leq \lambda_i^{''} \leq \lambda_i^{'} \leq \lambda_i \text{ and } \lambda_{i+2} \leq \lambda_i^{''} \leq \lambda_i.$$

In general: Let B be the principal submatrix of A of order $n-r$, then

$$\lambda_{i+r}(A) \leq \lambda_i(B) \leq \lambda_i(A), \quad i = 1, \cdots, n-r.$$

Theorem 6.1.9 (Perturbation theorem) *Let A, E be Hermitian. Then it holds*

$$\lambda_i(A) + \lambda_n(E) \leq \lambda_i(A+E) \leq \lambda_i(A) + \lambda_1(E), \quad i = 1, \cdots, n. \tag{1.9}$$

Proof: For $x \neq 0$, $R[x, A+E] = R[x, A] + R[x, E]$. Thus

$$R[x, A] + \lambda_n(E) \leq R[x, A+E] \leq R[x, A] + \lambda_1(E).$$

Applying (1.6) we get

$$\lambda_i(A) + \lambda_n(E) \leq \lambda_i(A+E) \leq \lambda_i(A) + \lambda_1(E).$$

■

Corollary 6.1.10 (Monotonic theorem) *If E is positive semidefinite, then it holds*

$$\lambda_i(A+E) \geq \lambda_i(A).$$

Corollary 6.1.11 (Weyl's theorem) *It holds*

$$\begin{aligned} & |\lambda_i(A+E) - \lambda_i(A)| \leq \max\{\lambda_1(E), -\lambda_n(E)\} \\ = \ & \max\{|\lambda_i(E)|, i = 1, \cdots, n\} = \rho(E) = \|E\|_2 \\ = \ & \textit{spectral radius of } E. \end{aligned}$$

Theorem 6.1.12 (Hoffmann-Wielandt) *If A, E are Hermitian matrices, then*

$$\sum_{i=1}^{n} (\lambda_i(A+E) - \lambda_i(A))^2 \leq \|E\|_F^2 = (\sum_{i=1}^{n} \lambda_i(E)^2)^{\frac{1}{2}}.$$

Proof: Later!

Definition 6.1.2 *A matrix $B = (b_{ij})$ is called double stochastic(d.s.), if (1) $b_{ij} \geq 0$. (2) $\sum_{j=1}^{n} b_{ij} = \sum_{j=1}^{n} b_{ji} = 1$, for $i, j = 1, \cdots, n$.*

Remark: The d.s. matrices form a convex set D.

Example: Let W be orthogonal and $W = (w_{ik})$. Then $(|w_{ik}|^2) = \tilde{W}$ is double stochastic.

Example: Let P be a permutation matrix. Then P is double stochastic (Extreme point of D).

Theorem 6.1.13 (Birkhoff) *D is the convex closure of the permutation matrices, that is, for $B \in D$, there exists $\alpha_1, \cdots, \alpha_r$ and $P_1, \cdots, P_r$ permutations such that*

$$B = \sum_{i=1}^{r} \alpha_i p_i, \quad \alpha_i \geq 0, \quad \sum_{i=1}^{r} \alpha_i = 1.$$

(Without Proof!)

Remark: Let l be a linear functional from D into R. Then it holds

$$\min_{P\in Perm.} l(P) \le l(B) = l(\sum_{i=1}^{r} \alpha_i p_i) \le \max_{P\in Perm.} l(P).$$

Proof of Hoffmann-Wielandt theorem 6.1.12: Let

$$A = U\Lambda U^*, \quad \Lambda = diag(\lambda_1(A), \cdots, \lambda_n(A)),$$

$$A + E = V\tilde{\Lambda}V^*, \tilde{\Lambda} = diag(\lambda_1(A+E), \cdots, \lambda_n(A+E)) \equiv diag(\tilde{\lambda}_i).$$

Then

$$\begin{aligned} -E &= A - (A+E) = U\Lambda U^* - V\tilde{\Lambda}V^* \\ &= V(V^*U\Lambda - \tilde{\Lambda}V^*U)U^* \\ &= V(W\Lambda - \tilde{\Lambda}W)U^* \end{aligned}$$

and since $W = V^*U$ is unitary, we have

$$\begin{aligned} \|E\|_F^2 &= \|W\Lambda - \tilde{\Lambda}W\|_F^2 = \sum_{i,k=1}^{n} |w_{ik}(\lambda_k - \tilde{\lambda}_i)|^2 \\ &= \sum_{i,k=1}^{n} |w_{ik}|^2 |\lambda_i - \tilde{\lambda}_k|^2 = l(\tilde{W}) \ge l(P) \text{ (for some } P) \\ &\quad \text{(Hereby } \tilde{W} = (|w_{ik}|) \text{ is in } D). \\ &= \sum_{k=1}^{n} |\lambda_k - \tilde{\lambda}_{\pi(k)}|^2 \text{ (for some permutation } \pi) \\ &= \min_{\pi} \sum_{k=1}^{n} |\lambda_k - \tilde{\lambda}_{\pi(k)}|^2 \\ &= \sum_{k=1}^{n} (\lambda_k(A) - \lambda_k(A+E))^2. \quad \text{(Exercise!)} \end{aligned}$$

■

Perturbation theorem of invariant subspaces (eigenvectors)

Theorem 6.1.14 *$A \in R^{n\times n}$ symmetric, $S \in R^{m\times m}$ symmetric and*

$$AQ_1 - Q_1 S = E_1 \ \textit{with} \ Q_1 \in R^{n\times m}, \ Q_1^T Q_1 = I_m. \tag{1.10}$$

Then there exist eigenvalues $\lambda_1', \cdots, \lambda_m'$ of A such that

$$|\lambda_i' - \lambda_i(S)| \le \|E_1\|_2. \tag{1.11}$$

Proof: Extend Q_1 to an orthogonal matrix $Q = (Q_1, Q_2)$, then

$$\begin{aligned} Q^T AQ &= \begin{pmatrix} Q_1^T AQ_1 & Q_1^T AQ_2 \\ Q_2^T AQ_1 & Q_2^T AQ_2 \end{pmatrix} = \begin{pmatrix} S + Q_1^T E_1 & E_1^T Q_2 \\ Q_2^T E_1 & Q_2^T AQ_2 \end{pmatrix} \quad \text{(by (1.10))} \\ &= \begin{pmatrix} S & 0 \\ 0 & Q_2^T AQ_2 - X \end{pmatrix} + \begin{pmatrix} Q_1^T E_1 & E_1^T Q_2 \\ Q_2^T E_1 & X \end{pmatrix} \equiv B + E. \end{aligned}$$

Here $Q_1^T E_1 = Q_1^T AQ_1 - S$ is symmetric. Corollary 6.1.10 results

$$|\lambda_i^{'} - \lambda_i(S)| \leq \|E\|_2.$$

Show that: $\|E\|_2 = \|E_1\|_2$ for suitable X.

It holds $\|E_1\|_2 \leq \|E\|_2$. The equality holds immediately from the *Extension Theorem* of *Kahan(1967)*:

Extension Lemma: Let $R = \begin{bmatrix} H \\ B \end{bmatrix}$, $H = H^*$. There exists a W such that the "*extend*" matrix $A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}$ satisfies $\|A\|_2 = \|R\|_2$.

Proof of Extension Lemma: Let $\rho = \|R\|_2$. For any choice of W we have $\rho^2 \leq \|A^2\|_2$ (by separation theorem). The theorem requires that for some W the matrix $\rho^2 - A^2$ is positive semidefinite.

Take any $\sigma > \rho$, show that $\sigma^2 - A^2 > 0$ for some W depending on σ. Then a limiting argument show that, as $\sigma \to \rho^+$, $\lim W(\sigma)$ exists.

For any W: Define $\tilde{R} = \begin{bmatrix} B^* \\ W \end{bmatrix}$. Write $A = (R, \tilde{R})$. Then

$$\sigma^2 - A^2 = \begin{bmatrix} I & 0 \\ L & I \end{bmatrix} \begin{bmatrix} \sigma^2 - R^*R & 0 \\ 0 & U(\sigma) \end{bmatrix} \begin{bmatrix} I & L^* \\ 0 & I \end{bmatrix}$$

and

$$\sigma^2 - RR^* = \begin{bmatrix} I & 0 \\ K & I \end{bmatrix} \begin{bmatrix} \sigma^2 - H^2 & 0 \\ 0 & V(\sigma) \end{bmatrix} \begin{bmatrix} I & K^* \\ 0 & I \end{bmatrix}.$$

where

$$\begin{aligned} U(\sigma) &= \sigma^2 - \tilde{R}^*[I + R(\sigma^2 - R^*R)^{-1}R^*]\tilde{R}, \\ V(\sigma) &= \sigma^2[I - B(\sigma^2 - H^2)^{-1}B^*]. \end{aligned}$$

Since $\sigma^2 > \rho^2 = \|R\|_2^2 = \|R^*R\|_2 = \|RR^*\|_2$, $\sigma^2 - R^*R$, $\sigma^2 - RR^*$ and $\sigma^2 - H^2$ are all positive definite. By Sylvester's Inertia theorem we have $V(\sigma)$ positive definite. $U(\sigma)$ depends on W.

The trick of the proof: To find a W such that $U(\sigma) = V(\sigma)$, and then from Sylvester's follows $\sigma^2 - A^2 > 0$.

First we prove that

$$W(\sigma) = -BH(\sigma^2 - H^2)^{-1}B^* = -B(\sigma^2 - H^2)^{-1}HB^*.$$

From above

$$U(\sigma) = \sigma^2 - BB^* - W^2 - (BH + WB)(\sigma^2 - H^2 - B^*B)^{-1}(HB^* + B^*W).$$

Consider

$$\begin{aligned} & (\sigma^2 - H^2 - B^*B)^{-1} \\ = \; & (\sigma^2 - H^2)^{-1} + (\sigma^2 - H^2)^{-1}B^*[I - B(\sigma^2 - H^2)^{-1}B^*]^{-1}B(\sigma^2 - H^2)^{-1} \\ & \quad \text{(Scherrman-Morrison formula)} \\ \equiv \; & S + SB^*XBS, \end{aligned}$$

where

$$S = (\sigma^2 - H^2)^{-1}$$

and

$$X = (I - B(\sigma^2 - H^2)^{-1}B^*)^{-1} = (I - BSB^*)^{-1}.$$

Set $Y = BSHB^*$. Then by $SH = HS$ we get

$$\begin{aligned} U(\sigma) &= \sigma^2 - BB^* - W^2 - (BH + HB)(S + SB^*XBS)(HB^* + B^*W) \\ &= \sigma^2 - BB^* - W^2 - BSH^2B^* + WY + YXY + WBSB^*XY + YW \\ &+ WBSB^*W + YXBSB^*W + WBSB^*W + WBSB^*XBSB^*W \\ &= V(\sigma) + \Omega \text{ (remainder term).} \end{aligned}$$

Then

$$\begin{aligned} \Omega &= W^2 + WY + YXY + W(I - X^{-1})XY + YW + W(I - X^{-1})W \\ &\quad + YX(I - X^{-1})W + W(I - X^{-1})X(I - X^{-1})W \\ &= YXY + WXY + YXW + WXW \\ &= (Y + W)X(Y + W) = 0. \end{aligned}$$

Thus

$$W(\sigma) = -Y = -BSHB^* = -B(\sigma^2 - H^2)^{-1}HB^*.$$

The matrix $W(\sigma)$ is a rational, and therefore meromorphic function of complex variable σ. Its only singularities are poles in any neighborhood of which $\|W\|_2$ must be unbounded. However $\|W\|_2 \leq \|A\|_2 < \sigma$ for all $\sigma > \rho$ and thus $W(\sigma)$ must be regular at $\sigma = \rho$ and so $W(\rho) = \lim_{\sigma \to \rho^+} W(\sigma)$. By continuity of norm we have

$$\|A(\rho)\|_2 = \lim_{\sigma \to \rho^+} \|A(\sigma)\|_2 = \rho.$$

■

Generalized Extension Theorem (C. Davis-Kahan-Weinberger)

Given H, B, E arbitary, then there exists W with

$$\left\| \begin{bmatrix} H & E \\ B & W \end{bmatrix} \right\|_2 = \max\{ \left\| \begin{bmatrix} H \\ B \end{bmatrix} \right\|_2, \; \|(H, E)\|_2 \}.$$

So for suitable X we have

$$\|E\|_2 = \| \begin{bmatrix} Q_1^T E_1 \\ Q_2^T E_1 \end{bmatrix} \|_2 = \|Q^T E_1\|_2 = \|E_1\|_2.$$

■

Theorem 6.1.15 *$A \in R^{n\times n}$ and $S \in R^{m\times m}$ are symmetric and $AX_1 - X_1S = E_1$, where $X_1 \in R^{n\times m}$ satisfies $\sigma_m(X_1) > 0$, then there exists eigenvalues $\lambda_1^{'}, \cdots, \lambda_m^{'}$ of A such that $|\lambda_i^{'} - \lambda_i(S)| \leq \|E_1\|_2/\sigma_m(X_1)$.*

Proof: Let $X_1 = Q_1R_1$ be the QR-decomposition of X_1. By substituting into $AX_1 - X_1S = E_1$ we get

$$AQ_1 - Q_1S = F_1,$$

where $S_1 = R_1 S {R_1}^{-1}$ and $F_1 = E_1 {R_1}^{-1}$. The theorem follows by applying theorem 6.1.14 and noting that $\lambda(S) = \lambda(S_1)$ and $\|F_1\|_2 \leq \|E_1\|_2/\sigma_m(X_1)$. ■

The eigenvalue bounds in theorem 6.1.14 depend on the size of the residual of the approximate invariant subspace, i.e., upon the size of $\|AQ_1 - Q_1S\|$. The following theorem tells how to choose S so that this quantity is minimized when $\|\cdot\| = \|\cdot\|_F$.

Theorem 6.1.16 *If $A \in R^{n\times n}$ is symmetric and $Q_1 \in R^{n\times m}$ satisfies $Q_1^T Q_1 = I_m$, then*

$$\min_{S\in R^{m\times m}} \|AQ_1 - Q_1S\|_F = \|AQ_1 - Q_1(Q_1^T AQ_1)\|_F = \|(I - Q_1Q_1^T)AQ_1\|_F.$$

Proof: Let $Q_2 \in R^{n\times(n-m)}$ be such that $Q = [Q_1, Q_2]$ is orthogonal. For any $S \in R^{m\times m}$ we have

$$\|AQ_1 - Q_1S\|_F^2 = \|Q^T AQ_1 - Q^T Q_1 S\|_F^2 = \|Q_1^T AQ_1 - S\|_F^2 + \|Q_2^T AQ_1\|_F^2.$$

Clearly, the minimizing S is given by $S = {Q_1}^T AQ_1$. ■

Theorem 6.1.17 *Suppose $A \in R^{n\times n}$ is symmetric and $Q_1 \in R^{n\times k}$ satisfies $Q_1^T Q_1 = I_k$. If*

$$Z^T(Q_1^T AQ_1)Z = diag\,(\theta_1, \cdots \theta_k) = D$$

is the Schur decomposition of $Q_1^T AQ_1$ and $Q_1Z = [y_1, \cdots, y_k]$, then

$$\|Ay_i - \theta_i y_i\|_2 = \ |(I - Q_1Q_1^T)AQ_1Ze_i\|_2 \leq \|(I - Q_1Q_1^T)AQ_1\|_2$$

for $i = 1, \cdots, k$. The θ_i are called Ritz values*, the y_i are called* Ritz vectors*, and the (θ_i, y_i) are called* Ritz pairs*.*

Proof: $Ay_i - \theta_i y_i = AQ_1Ze_i - Q_1ZDe_i = (AQ_1 - Q_1(Q_1^T AQ_1))Ze_i$. The theorem follows by taking norms. ■

Definition 6.1.3 *The inertia of a symmetric matrix A is a triplet of integers (m, z, p), where m, z and p are the number of negative, zero and positive elements of $\sigma(A)$.*

Theorem 6.1.18 (Sylvester Law of Interia) *If $A \in R^{n\times n}$ is symmetric and $X \in R^{n\times n}$ in nonsingular, then A and X^TAX have the same inertia.*

Proof: Suppose $\lambda_r(A) > 0$ and define the subspace $S_0 \subset R^n$ by

$$S_0 \;=\; Span\{X^{-1}q_1, \cdots, X^{-1}q_r\}, \;\; q_i \neq 0,$$

where $Aq_i = \lambda_i(A)q_i$ and $i = 1, \cdots, r$. From the Minimax characterization of $\lambda_r(X^TAX)$ we have

$$\lambda_r(X^TAX) = \max_{dim(S)=r} \min_{y\in S} \frac{y^T(X^TAX)y}{y^Ty} \geq \min_{y\in S_0} \frac{y^T(X^TAX)y}{y^Ty}.$$

Now for any $y \in R^n$ we have

$$\frac{y^T(X^TX)y}{y^Ty} \geq \sigma_n(X)^2, \text{ while for } y \in S_0.$$

It is clear that

$$\frac{y^T(X^TAX)y}{y^T(X^TX)y} \geq \lambda_r(A).$$

Thus,

$$\lambda_r(X^TAX) \geq \min_{y\in S_0}\{\frac{y^T(X^TAX)y}{y^T(X^TX)y}\frac{y^T(X^TX)y}{y^Ty}\} \geq \lambda_r(A)\sigma_n(X)^2.$$

An analogous argument with the roles of A and X^TAX reversed shows that

$$\lambda_r(A) \geq \lambda_r(X^TAX)\sigma_n(X^{-1})^2 = \lambda_r(X^TAX)/\sigma_1(X)^2.$$

It follows that A and X^TAX have the same number of positive eigenvalues. If we apply this result to $-A$, we conclude that A and X^TAX have the same number of negative eigenvalues. Obviously, the number of zero eigenvalues possessed by each matrix is also the same. ■

6.2 Tridiagonalization and the Symmetric QR-algorithm

We now investigate how the practical QR algorithm develop in Chapter 1 can be specialized when $A \in R^{n\times n}$ is symmetric. There are three obvious observations:

(a) If ${Q_0}^T A Q_0 = H$ is upper Hessenberg, then $H = H^T$ must be tridiagonal.

(b) Symmetry and tridiagonal band structure are preserved when a single shift QR step is performed.

(c) There is no need to consider complex shift, since $\sigma(A) \subset R$.

Algorithm 2.1 (Householder Tridiagonalization) Given symmetric $A \in R^{n\times n}$, the following algorithm overwrites A with ${Q_0}^T A Q_0 = T$, where T is tridiagonal and $Q_0 = P_1 \cdots P_{n-2}$ is the product of Householder transformations.

For $k = 1, 2, \cdots, n-2$,
determine a Householder $P_k \in R^{n-k}$ such that

$$\bar{P}_k \begin{bmatrix} a_{k+1,k} \\ \vdots \\ a_{nk} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$A := P_k A P_k^T, P_k = diag(I_k, \bar{P}_k)$.

This algorithm requires $\frac{2}{3}n^3$ flops. If Q_0 is required, it can be formed with an additional $(2/3)n^3$ flops.

We now consider the single shift QR iteration for symmetric matrices.

$$\begin{array}{l} T = Q_0^T A Q_0, \quad \text{(tridiagonal)} \\ \text{For } k = 0, 1, \cdots \\ T - \mu I = QR, \quad \text{(QR decomposition)} \\ T := RQ + \mu I. \end{array} \tag{6.2.1}$$

Single Shift: Denote T by

$$T = \begin{bmatrix} a_1 & b_2 & & \\ b_2 & a_2 & \ddots & \\ & \ddots & \ddots & b_n \\ & & b_n & a_n \end{bmatrix}.$$

We can set **(a)** $\mu = a_n$ or **(b)** a more effective choice to shift by the eigenvalues of $\begin{bmatrix} a_{n-1} & b_n \\ b_n & a_n \end{bmatrix}$ that is closer to a_n. This is known as the Wilkinson shift and is given by

$$\begin{array}{l} \mu = a_n + d - sign(d)\sqrt{d^2 + b_n^2}, \quad \text{where} \\ d = (a_{n-1} - a_n)/2. \end{array} \tag{6.2.2}$$

Wilkinson (1968) has shown that (6.2.2) is cubically convergent with either shift strategy, but gives heuristic reasons why (6.2.2) is prefered.

Implicit Shift:

As in the unsymmetric QR iteration, it is possible to shift implicitly in (6.2.1). Let $c = \cos(\theta)$ and $s = \sin(\theta)$ by computed such that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^T \begin{bmatrix} a_1 - \mu \\ b_2 \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$$

then $J_1 e_1 = Q e_1$, where $Q^T T Q = RQ + \mu I = \bar{T}$ (as in (6.2.1)) and $J_1 = J(1, 2, \theta)$.

$$J_1{}^T T J_1 = \begin{bmatrix} \times & \times & + & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ + & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix}.$$

We are thus in a position to apply implicit Q theorem provided we can compute rotations $J_2, \cdots, J_{n-1}$ with the property that if $Z = J_1 \cdots J_{n-1}$ then $Ze_1 = J_1 e_1 = Q e_1$ and $Z^T T Z$ is tridiagonal.

$$T := J_2^T T J_2 = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & + & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & + & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix}, \quad T := J_3^T T J_3 = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & + \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & + & \times & \times \end{bmatrix}$$

$$T := J_4^T T J_4 = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix}$$

Algorithm 2.2 (Implicit Symmetric QR step with Wilkinson Shift) Given an unreduced symmetric tridiagonal matrix $T \in R^{n \times n}$, the following algorithm overwrites T with $\bar{Z}^T T \bar{Z}$, where $\bar{Z} = J_1 \cdots J_{n-1}$ is the product of Givens rotation with $\bar{Z}^T(T - \mu I)$ is upper triangular and μ is Wilkinson shift.

$d := (t_{n-1,n-1} - t_{nn})/2,$
$\mu = t_{nn} - t_{n,n-1}^2 / [d + sign(d)\sqrt{d^2 + t_{n,n-1}^2}],$
$x := t_{11} - \mu,$
$z := t_{21},$
For $k = 1, \cdots, n-1,$
determine $c = \cos(\theta), s = \sin(\theta)$
such that $\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$
$T := J_k^T T J_k, \;\; J_k = J(k, k+1, \theta).$
If $k < n-1,$ then $x := t_{k+1,k}, \; z = t_{k+2,k}.$

Algorithm 2.3 (Symmetric QR algorithm) Given symmetric matrix $A \in R^{n\times n}$ and a tolerance ϵ, the following algorithm overwrites A with $Q^TAQ = D + E$, where Q is orthogonal, D is diagonal and E satifies $\|E\|_2 \simeq eps\|A\|_2$.

Using Algorithm 2.1 compute
$A := (P_1 \cdots P_{n-1})^T A(P_1 \cdots P_{n-2}) = T$.
Repeat set $a_{i+1,i}$ and $a_{i,i+1}$ to zero if
$|a_{i+1,i}| = |a_{i,i+1}| \leq \epsilon(|a_{ii}| + |a_{i+1,i+1}|)$
for any $i = 1, \cdots, n-1$.
Find the largest q and the smallest p such that if

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix} \begin{matrix} \}p \\ \}n-p-q \\ \}q \end{matrix}$$

then A_{33} is diagonal and A_{22} has no zero subdiagonal elements.
If $q = n$ then stop.
Apply algorithm 2.2 to A_{22}, $A = diag(I_p, \bar{Z}, I_q)^T\ A\ diag(I_p, \bar{Z}, I_q)$,
Go to Repeat.

This algorithm requires about $(2/3)n^3$ flops and about $5n^3$ flops if Q is accumulated.

6.3 Once Again:The Singular Value Decomposition

Let $A \in R^{m\times n}$. If $U^TAV = diag(\sigma_1, \cdots, \sigma_n)$ is the SVD of A $(m \geq n)$ then

$$V^T(A^TA)V = diag(\sigma_1^2, \cdots, \sigma_n^2) \in R^{n\times n} \text{ and} \tag{3.1}$$
$$U^T(AA^T)U = diag(\sigma_1^2, \cdots, \sigma_n^2, 0, \cdots, 0) \in R^{n\times n}. \tag{3.2}$$

Moreover if $U = \left[\ U_1,\ \ U_2\ \right]$ and we define the orthogonal Q by

$$Q = \frac{1}{\sqrt{2}}\begin{bmatrix} V & V & 0 \\ U_1 & -U_1 & \sqrt{2}\,U_2 \end{bmatrix},$$

then

$$Q^T\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}Q = diag(\sigma_1, \cdots, \sigma_n, -\sigma_1, \cdots, -\sigma_n, 0, \cdots, 0). \tag{3.3}$$

These connections to the symmetric eigenvalue problem allow us to develop an algorithm for SVD as previous section.

Theorem 6.3.1 *If $A \in R^{m\times n}$, then for $k = 1, \cdots, \min\{m, n\}$,*

$$\sigma_k(A) = \max_{dimS=k, dimT=k}\{\min_{x\in S, y\in T}\frac{y^TAx}{\|x\|_2\|y\|_2}\} = \max_{dimS=k}\{\min_{x\in S}\frac{\|Ax\|_2}{\|x\|_2}\}.$$

Proof: Exercise! Prove theorem 6.1.5 (Weyl) and theorem 6.1.6 (Courant-Fisher)! ■

By applying theorem 6.1.9 to $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & (A+E)^T \\ (A+E) & 0 \end{bmatrix}$ and theorem6.1.8 to A^TA we obtain

Corollary 6.3.2 *If A and $A + E$ are in $R^{m\times n}(m \geq n)$, then for $k = 1, 2, \cdots, n$*

$$|\sigma_k(A + E) - \sigma_k(A)| \leq \sigma_1(E) = \|E\|_2.$$

Corollary 6.3.3 *Let $A = [a_1, \cdots, a_n]$ be a column partitioning of $A \in \mathbf{R}^{m\times n}(m \geq n)$. If $A_r = [a_1, \cdots, a_r]$, then for $r = 1, \cdots, n-1$,*

$$\sigma_1(A_{r+1}) \geq \sigma_1(A_r) \geq \sigma_2(A_{r+1}) \geq \cdots \geq \sigma_r(A_{r+1}) \geq \sigma_r(A_r) \geq \sigma_{r+1}(A_{r+1}).$$

Theorem 6.3.4 *If A and $A + E$ are in $\mathbf{R}^{m\times n}(m \geq n)$, then*

$$\sum_{k=1}^{n}[\sigma_k(A + E) - \sigma_k(A)]^2 \leq \|E\|_F^2.$$

Proof: Apply Theorem 6.1.12 to $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & (A+E)^T \\ (A+E) & 0 \end{bmatrix}$. ■

We now show a variant of the QR algorithm can be used to comput SVD of a matrix. Equation (3.1) suggests:

(a) Form $C = A^T A$;

(b) Use the symmetric QR algorithm to compute $V_1^T C V_1 = diag(\sigma^2)$;

(c) Use QR with column pivoting to upper triangularize $B = AV_1$:

$$U^T(AV_1)\Pi = R.$$

Since R has orthogonal columns, it follows that $U^T A(V_1\Pi)$ is diagonal.

A preferable method for computing the SVD is described in Golub and Kahan(1965). The first step is to reduce A to upper bidiagonal form using algorithm 7.5 or 7.6 in part I:

$$U_B^T A V_B = \begin{bmatrix} B \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} d_1 & f_2 & & \bigcirc \\ & d_2 & \ddots & \\ & & \ddots & f_n \\ \bigcirc & & & d_n \\ \cdots & \cdots & \cdots & \cdots \\ & \bigcirc & & \end{bmatrix}.$$

The remaining problem is thus to compute the SVD of B. Consider applying an implicit QR step (algorithm 8.2) to the tridiagonal matrix $T = B^T B$:

(a) Compute the eigenvalue λ of $\begin{bmatrix} d_m^2 + f_m^2 & d_m f_n \\ d_m f_n & d_n^2 + f_n^2 \end{bmatrix}$ $(m = n-1)$ that is closer to $d_n^2 + f_n^2$.

(b) Compute $c_1 = \cos\theta_1$ and $s_1 = \sin\theta_1$ such that

$$\begin{bmatrix} c_1 & -s_1 \\ s_1 & c_1 \end{bmatrix} \begin{bmatrix} d_1^2 - \lambda \\ d_1 f_2 \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix},$$

and set $J_1 = J(1, 2, \theta_1)$.

(c) Compute Givens rotations $J_2, \cdots, J_{n-1}$ such that if $Q = J_1 \cdots J_{n-1}$ then $Q^T T Q$ is tridiagonal and $Qe_1 = J_1 e_1$.

Note that these calculations require the explicit formation of $B^T B$, which is unwise in the numerical standpoint. Suppose instead that we apply Givens rotation J_1 above to B directly. This gives

$$B := BJ_1 = \begin{bmatrix} \times & \times & & & \\ + & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{bmatrix}, \quad n = 5.$$

Determine Givens rotations $U_1, V_2, U_2, \cdots, V_{n-1}$ and U_{n-1} to chase the nonzero element down the diagonal:

$$B := U_1^T B = \begin{bmatrix} \times & \times & + & & \\ & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{bmatrix}, \quad B := BV_2 = \begin{bmatrix} \times & \times & & & \\ & \times & \times & & \\ & + & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{bmatrix},$$

$$B := U_2^T B = \begin{bmatrix} \times & \times & & & \\ & \times & \times & + & \\ & & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{bmatrix}, \quad B := BV_3 = \begin{bmatrix} \times & \times & & & \\ & \times & \times & & \\ & & \times & \times & \\ & & + & \times & \times \\ & & & & \times \end{bmatrix}$$

The process terminates with a new bidiagonal $\bar{B}$ as follows

$$\bar{B} = (U_{n-1}^T \cdots U_1^T) B (J_1 V_2 \cdots V_{n-1}) = \bar{U}^T B \bar{V}.$$

Since each V_i has the form $V_i = J(i, i+1, \theta_i),\ i = 2, \cdots, n-1$, it follows that $\bar{V} e_1 = Q e_1$. By implicit Q theorem we can assert that $\bar{V}$ and Q are essentially the same. Thus we can implicitly effect the transition from T to $\bar{T} = \bar{B}\bar{B}^T$ by working directly on the bidiagonal matrix.

It is necessary for these claims to hold that the underlying tridiagonal matrices be unreduced. This is the condition for the performance of implicit QR method.

Let $B = \begin{bmatrix} d_1 & f_2 & & \bigcirc \\ & d_2 & \ddots & \\ & & \ddots & f_n \\ \bigcirc & & & d_n \end{bmatrix}$. If $(B^T B)_{i,i+1} = f_{i+1} d_i = 0$, then:

Either $f_{i+1} = 0$: B is reduced to $B = \left(\begin{array}{c|c} B_1 & \bigcirc \\ \hline \bigcirc & B_2 \end{array}\right)$ two small problems.

Or d_i=0: What happens?

For Example

$$B = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & 0 \\ 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{bmatrix}, \ (d_2 = 0, n = 5)$$

$$\begin{array}{c} \longrightarrow \\ \text{Rotation} \\ \text{in } (2,3) \end{array} \quad B := J_1(2,3,\theta)B = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 \\ 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{bmatrix}$$

$$\begin{array}{c} \longrightarrow \\ \text{Rotation} \\ \text{in } (2,4) \end{array} \quad B := J_2(2,4,\theta)B = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \times \\ 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{bmatrix}$$

$$\begin{array}{c} \longrightarrow \\ \text{Rotation} \\ \text{in}(2,5) \end{array} \quad B := J_3(2,5,\theta)B = \left[\begin{array}{cc|ccc} \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{array}\right]$$

Criteria: For smallness within $B's$ band are usually of the the form

$$|f_i| \leq \epsilon(|d_{i-1}| + |d_i|) \quad \text{and} \quad |d_i| \leq \epsilon\|B\|,$$

where ϵ is a small multiple of the unit roundoff.

Algorithm 3.1 (Golub-Kahan SVD Step) $B \in R^{n\times n}$ is bidiagonal having nonzero subdiagonal and diagonal, the following algorithm overwrites B with the bidiagonal matrix $\bar{B} = \bar{U}^T B \bar{V}$, where $\bar{U}$ and $\bar{V}$ are orthogonal and $\bar{V}$ is essentially the orthogonal matrix that would be obtained by applying algorithm 8.2 to $T = B^T B$. Let μ be the eigenvalue of the trailing 2×2 sumatrix of $T = B^T B$ that is closer to t_{nn}.

$y = t_{11} - \mu$
$z = t_{12}$
For $k = 1, \cdots, n-1$,
  Determine $c = \cos\theta$ and $s = \sin\theta$ such that
  $[y \ , \ z] \begin{bmatrix} c & s \\ -s & c \end{bmatrix} = [* \ , \ 0]$
  $B = BJ(k, k+1, \theta)$
  $y = b_{kk}$
  $z = b_{k+1,k}$
  Determine $c = \cos\theta$ and $s = \sin\theta$ such that
  $\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}$
  $B := J(k, k+1)^T B$
  If $k < n-1$, then $y := b_{k,k+1}, \ z := b_{k,k+2}$.

This algorithm requires $20n$ flops and $2n$ square roots. Accumulating $\bar{U}$ requires $4mn$ flops and $\bar{V}$ requires $4n^2$ flops.

Algorithm 3.2 (The SVD Algorithm) Given $A \in R^{m\times n}\,(m \geq n)$ and ϵ a tolerance, the following algorithm overwrites A with $U^T A V = D + E$, where $U \in R^{m\times n}$ is orthogonal, $V \in R^{n\times n}$ is orthogonal, $D \in R^{m\times n}$ is diagonal, and E satisfies $\|E\|_2 \simeq eps\|A\|_2$. Using algorithm 7.5 or 7.6 in Part I to compute the bidiagonalization $A := (U_1 \cdots U_n)^T A (V_1 \cdots V_{n-2})$.

Repeat

Set $a_{i,i+1}$ to zero if $|a_{i,i+1}| \leq \epsilon(|a_{ii}| + |a_{i+1,i+1}|)$ for any $i = 1, \cdots, n-1$.

Find the largest q and the smallest p such that

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} p \\ n-p-q \\ q \\ m-n \end{matrix}$$

Then A_{33} is diagonal and A_{22} has a nonzero subdiagonal.

If $q = n$ then stop.

If any diagonal entry in A_{22} is zero then zero the subdiagonal entry in the same row and go to **Repeat**.

Apply algorithm 3.1 to A_{22},

$A := diag(I_p, \bar{U}, I_{q+m-n})^T A\, diag(I_p, \bar{V}, I_q)$.

Go to Repeat

6.4 Jacobi Methods

Jacobi(1846) proposed a method for reducing a Hermitian matrix $A = A^* \in \mathbf{C}^{n\times n}$ to diagonal form using Givens rotations. Let $A \in \mathbf{C}^{n\times n}$ be a Hermitian matrix, there exists a unitary U such that

$$U^*AU = diag(\lambda_1, \cdots, \lambda_n). \tag{4.1}$$

The Jacobi method constructs U as the product of infinite many two dimensional Givens rotations. Fix indices i, k, $i \neq k$, Given a Givens Rotation

$$U_{ik} = \begin{pmatrix} 1 & & 0 & \vdots & & & & \vdots & & & & \\ & \ddots & & \vdots & & \bigcirc & & \vdots & & \bigcirc & & \\ 0 & & 1 & \vdots & & & & \vdots & & & & \\ \cdots & \cdots & \cdots & e^{i\alpha}cos\varphi & \cdots & \cdots & \cdots & sin\varphi & \cdots & \cdots & \cdots & i \\ & & & \vdots & 1 & & 0 & \vdots & & & & \\ & \bigcirc & & \vdots & & \ddots & & \vdots & & \bigcirc & & \\ & & & \vdots & 0 & & 1 & \vdots & & & & \\ \cdots & \cdots & \cdots & -sin\varphi & \cdots & \cdots & \cdots & e^{i\alpha}cos\varphi & \cdots & \cdots & \cdots & k \\ & & & \vdots & & & & \vdots & 1 & & 0 & \\ & \bigcirc & & \vdots & & \bigcirc & & \vdots & & \ddots & & \\ & & & \vdots & & & & \vdots & 0 & & 1 & \\ & & & i & & & & k & & & & \end{pmatrix}. \tag{4.2}$$

Hereby α, φ are free parameters, if A is real symmetric then $\alpha = 0$. Set $V = U_{ik}$, $B = V^*AV$. Then

$$b_{sj} = \left\{ \begin{array}{ll} a_{sj}, & s \neq i,k \\ e^{i\alpha}cos\varphi a_{ij} - sin\varphi a_{kj}, & s = i \\ sin\varphi a_{ij} + e^{i\alpha}cos\varphi a_{kj}, & s = k \end{array} \right\} \; j \neq i,k \tag{4.3}$$

$$\left\{ \begin{array}{ll} b_{si} = e^{i\alpha}cos\varphi a_{si} - sin\varphi a_{sk}, & s \neq i,k \\ b_{sk} = sin\varphi a_{si} + e^{i\alpha}cos\varphi a_{sk}, & s \neq i,k \end{array} \right. \tag{4.4}$$

$$\left\{ \begin{array}{l} b_{ik} = sin\varphi cos\varphi e^{i\alpha}(a_{ii} - a_{kk}) + e^{2i\alpha}(cos^2\varphi - sin^2\varphi)a_{ki} \\ b_{ki} = \overline{b}_{ik} \\ b_{ii} = cos^2\varphi a_{ii} + sin^2\varphi a_{kk} - sin\varphi cos\varphi[e^{-i\alpha}a_{ki} + e^{i\alpha}a_{ik}] \\ b_{kk} = sin^2\varphi a_{ii} + cos^2\varphi a_{kk} - sin\varphi cos\varphi[e^{-i\alpha}a_{ki} + e^{i\alpha}a_{ik}] \end{array} \right. \tag{4.5}$$

We denote here the Frobenius norm (Hilbert-Schmidt norm) by $\epsilon(A) = \sqrt{\sum_{i,k} |a_{ik}|^2}$ and define the "outer norm" by

$$g(A) = \sqrt{\sum_{i\neq k} |a_{ik}|^2},$$

which is only a seminorm (that is $g(A) = 0 \Rightarrow A = 0$ does not hold). We also have $\epsilon(UA) = \epsilon(A) = \epsilon(AV)$ for unitary U, V. Therefore

$$\epsilon(A) = \epsilon(V^*AV) = \epsilon(B),$$

that is

$$\sum_{j,s} |a_{js}|^2 = \sum_{j,s} |b_{js}|^2. \tag{4.6}$$

On the other hand one computes

$$|a_{ii}|^2 + |a_{kk}|^2 + 2|a_{ik}|^2 = |b_{ii}|^2 + |b_{kk}|^2 + 2|b_{ik}|^2. \tag{4.7}$$

Together with $b_{jj} = a_{jj}$ $j \neq i, k$ follows from (4.6) and (4.7)

$$\sum_{j \neq s} |a_{js}|^2 = \sum_{j \neq s} |b_{js}|^2 + 2|a_{ik}|^2 - 2|b_{ik}|^2$$

or

$$g^2(A) - 2|a_{ik}|^2 + 2|b_{ik}|^2 = g^2(B). \tag{4.8}$$

To make $g(B)$ as small as possible we choose the free parameters α, φ such that $b_{ik} = 0$. Multiplying the first equation in (4.5) by $2e^{-i\alpha}$ and set b_{ik} to zero:

$$sin2\varphi(a_{kk} - a_{ii}) = 2e^{i\alpha}cos^2\varphi a_{ik} - 2e^{-i\alpha}sin^2\varphi \bar{a}_{ik}. \tag{4.9}$$

We exclude the trivial case $a_{ik} = 0$ (then set $V = I$). Suppose that $a_{ik} \neq 0$. Compare the imaginary part in (4.9) which results $0 = Im(a_{ik}e^{i\alpha})$. This equation holds for $\alpha = -\text{arg}a_{ik}$. From $a_{ik}e^{i\alpha} = |a_{ik}|$. (4.9) leads to

$$2|a_{ik}|(cos^2\phi - sin^2\phi) = sin2\phi(a_{kk} - a_{ii}),$$

where

$$cot2\phi = \frac{a_{kk} - a_{ii}}{2|a_{ik}|}. \tag{4.10}$$

(4.10) has exactly one solution in $(-\frac{\pi}{4}, \frac{\pi}{4}]$. The choice $\alpha = -arga_{ik} + \pi$ leads to the same matrix B. For symmetric A, we choose $\alpha = 0$, then ϕ is obtained by

$$cot2\phi = \frac{a_{kk} - a_{ii}}{2a_{ik}}.$$

So the Jacobi method proceeds: $A_0 := A$, an iteration sequence $A_0, A_1, \cdots$ is constructed by $A_{m+1} := V_m^* A_m V_m$, $A_m = (a_{ik}^m)$. Hereby V_m has the form of (4.2). The underlying pivot pairs i, k of V_m is formed according to a rule of choice so that the underlying α, ϕ are chosen satisfying $a_{ik}^{m+1} = 0$.

Choice rules:
(1) choose (i, k) such that

$$|a_{ik}^m| = \max_{j \neq s} |a_{js}^m|.$$

This is the **classical Jacobi method**.

Theorem 6.4.1 *Let A be Hermitian. $V = U_{ik}$ is as in (4.2) where (i,k) are chosen so that $|a_{ik}|$ is maximal with α, ϕ according to (4.10). Let $B = V^*AV$. Then it holds*

$$g^2(B) \le p^2 g^2(A) \quad \text{with} \quad p = \sqrt{\frac{n^2-n-2}{n^2-n}} < 1. \tag{4.11}$$

Proof: There are n^2-n off-diagonal elements, so $g^2(A) \le (n^2-n)|a_{ik}|^2$. Thus $|a_{ik}|^2 \ge \frac{1}{n^2-n} g^2(A)$, hence

$$g^2(B) = g^2(A) - 2|a_{ik}|^2 \le \frac{n^2-n-2}{n^2-n} g^2(A).$$

■

Theorem 6.4.2 *The classcial Jacobi method converges, that is, there exists a diagonal matrix Λ so that $\lim_{m\to\infty} A_m = \Lambda$.*

Proof: From (4.11) follows $g(A_m) \to 0$, so $a_{rs}^{(m)} \to 0$ for all $r \ne s$. It remains to show the convergence of diagonal elements. From (4.5) and (4.10) follows that

$$\begin{aligned} |b_{ii} - a_{ii}| &= |\sin^2\phi(a_{kk} - a_{ii}) - |a_{ik}| 2\sin\phi\cos\phi| \\ &= |a_{ik}|\ |2\sin^2\phi\cot 2\phi - 2\sin\phi\cos\phi| \\ &= |a_{ik}|\ |\frac{\sin\phi}{\cos\phi}| \le |a_{ik}|. \end{aligned}$$

Analogously, $|b_{kk} - a_{kk}| \le |a_{ik}|$. If now i,k are the pivot indices of A_m, then from above we have

$$|a_{jj}^m - a_{jj}^{m+1}| \le |a_{ik}^m| \le g(A_m) \le p^m g(A).$$

Thus

$$|a_{jj}^{m+q} - a_{jj}^m| \le |p^m + p^{m+1} + \cdots + p^{m+q-1}| g(A) \le \frac{p^m}{1-p} g(A).$$

This shows that the convergence of diagonal. ■

Schonage (1964) and Van Kempen (1966) show that for k large enough there is a constant c such that $g(A_{k+N}) \le c g(A_k)^2$, $N = \frac{n(n-1)}{2}$, i.e., quadratic convergence. An earlier result established by Henrici (1958) when A has distinct eigenvalue.

(2) choose (i,k) cyclically, e.g., $(i,k) = (1,2),(1,3),\ldots,(1,n);\ (2,3),\ldots,(2,n);\ \ldots;(n-1,n);(1,2),(1,3),\ldots$. This is the **cyclic Jacobi method**.

Algorithm 4.1 (Serial Jacobi cyclic Jacobi) Given a symmetric $A \in \mathbf{R}^{n\times n}$ and $\delta \ge eps$, the following algorithm overwrites A with $U^TAU = D + E$, where U is orthogonal, D is diagonal, and E has a zero diagonal and satisfies $\| E \|_F \le \delta \| A \|_F$:

$\delta := \delta \| A \|_F$
Do until $g(A) \le \delta^2$
For $p = 1, 2, \ldots, n-1$,
  For $q = p+1, \cdots, n$,
    Find $J = J(p,q,\theta)$ such that the (p,q) entry of J^TAJ is zero,
  $A := J^TAJ$.

This algorithm requires $2n^3$ flop per sweep. An additional $2n^3$ flop are required if U is accumulated. (Hereby it is customary to refer to each set of $\frac{n(n-1)}{2}$ rotations as a sweep).

A proof of quadratic convergence see Wilkinson (1962) and Van kempen (1966).

Remark 4.1 In classical Jacobi method for each update $O(n^2)$ comparsions are required in order to locate the largest off-diagonal element. Thus much more time is spent by searching than updating. So the cyclic Jacobi method is considerably faster than classical Jacobi method.

(3) When implementing serical Jacobi method, it is sensible to skip the annihilation of a_{ik} if its modulus is less than some small (sweep-dependent) parameter, because the net reduction of $g(A)$ is not worth to cost. This leads to what is called **threshold Jacobi method**.

Given a threshold value δ, choose the indices pair (i,k) as in (2). But perform the rotation only for $|a_{ik}^m| > \delta$. If all $|a_{ik}^m| \leq \delta$, then we substitute δ by $\delta/2$ and so on. Details concering this variant of Jacobi's algorithm may be found in Wilkinson (AEP p.277ff).

Remark 4.2 (1) Although the serial Jacobi method (2) and (3) converge quadratically, it is not competitive with symmetric QR algorithm. One sweep of Jacobi requires as many flops as a complete computation of symmetric QR algorithm. However, the Jacobi iteration is attractive, for example, the matrix A might be close to a diagonal form. In this situation, the QR algorithm loses its advantage.
(2) The Jacobi iteration is adapted to parallel computation. A given computational task, such as a sweep, can be shared among the various CPUs thereby reducing the overall computation time.
(3) In practice we usually apply the choice (2) or (3).
(4) It is not necessary to determine ϕ explicitly in (4.10), since only $c = \cos\phi$ and $s = \sin\phi$ are needed. From (4.10) follows $\frac{1-4s^2+4s^4}{s^2(1-s^2)} = \frac{(a_{kk}-a_{ii})^2}{4|a_{ik}|^2}$, a quadratic equation in s^2. The sign is determined by (4.10).

6.5 Some Special Methods

6.5.1 Bisection method for tridiagonal symmetric matrices

Let A be tridiagonal, real and symmetric. Write

$$A = \begin{bmatrix} a_1 & b_1 & 0 & \dots & & 0 \\ b_1 & a_2 & b_2 & & & \vdots \\ 0 & b_2 & a_3 & & & \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ & & & & \ddots & b_{n-1} \\ 0 & \dots & & 0 & b_{n-1} & a_n \end{bmatrix}. \tag{5.1}$$

Let A_k be the kth principal submatrix

$$A_k = \begin{bmatrix} a_1 & b_1 & 0 & \dots & & 0 \\ b_1 & a_2 & b_2 & & & \vdots \\ 0 & b_2 & a_3 & & & \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ & & & & \ddots & b_{k-1} \\ 0 & \dots & & 0 & b_{k-1} & a_k \end{bmatrix}$$

and

$$f_k(\lambda) = \det(\lambda I_k - A_k), \text{ for } k = 1, \cdots, n. \tag{5.2}$$

$$(f_n(\lambda) = \text{ Characteristic polynomial of } A.)$$

Write $f_0(\lambda) = 1$ and $f_1(\lambda) = \lambda - a_1$ we have the recursive formula:

$$f_k(\lambda) = (\lambda - a_k) f_{k-1}(\lambda) - b_{k-1}^2 f_{k-2}(\lambda), \ k = 2, \dots, n. \tag{5.3}$$

It holds:

Theorem 6.5.1 *If $b_i \neq 0$ in (5.1) for $i = 1, \dots, n$, then $f_k(\lambda)$ has k real simple roots, $k = 0, \dots, n$. For $1 \leq k \leq n-1$ the roots of $f_k(\lambda)$ separate the roots of $f_{k+1}(\lambda)$.*

Proof: Since A_k is real symmetric, it follows form (5.2) that the roots of $f_k(\lambda)$ are real. The rank of $\lambda I_k - A_k$ is at least $k-1$ (scratsch the first row and k-th column, and then consider $b_i \neq 0$), therefore the dimension of the zero spaces of $\lambda I_k - A_k$ is not bigger than one, so we have simple roots.

$n = 2 : f_1$ has the root a_1 and $f_2(a_1) = -b_1^2 < 0$ (from (5.3), $k = 2$ and $\lambda = a_1$), both roots of f_2 must lie on the right and left sides of a_1, respectively.

Suppose the assertion is true for $k = 2, \dots, n-1$, we shall prove that it is also true for $k = n$. It only needs to show that the roots of f_{n-1} separate the roots of f_n.

Let $\mu_1 > \mu_2 > \cdots > \mu_{n-1}$ be the roots of f_{n-1}. From (5.3) we have

$$\begin{aligned} f_n(\mu_i) &= -b_{n-1}^2 f_{n-2}(\mu_i), \\ f_n(\mu_{i+1}) &= -b_{n-1}^2 f_{n-2}(\mu_{i+1}). \end{aligned} \tag{5.4}$$

The roots of f_{n-2} separate the roots of f_{n-1}, there exists exactly one root of f_{n-2} between μ_i and μ_{i+1}, that is, $\mathrm{sgn} f_{n-2}(\mu_i) = -\mathrm{sgn} f_{n-2}(\mu_{i+1})$. Therefore it holds also for f_n (because of (5.4)), so there is at least one root of f_n in $(\mu_{i+1},\ \mu_i)$ from Roll'e theorem, for $i = 1, \ldots, n-2$. It is $f_n(\mu_1) = -b_{n-1}^2 f_{n-2}(\mu_1) < 0$, since $f_{n-2}(\lambda) = \lambda^{n-2} + \cdots$ and all roots of f_{n-2} are on the left side of μ_1.

On the other hand $f_n \to \infty$ for $\lambda \to \infty$, so there exists an other root of f_n in $(\mu_1,\ \infty)$. Similarly, we can show that there is a root of f_n in $(-\infty,\ \mu_{n-1})$. This shows that f_n has n distinct, simple roots, which are separated by the roots of f_{n-1}. ■

The sequence of functions $f_0, f_1, \cdots, f_n$ satisfies in each bounded interval $[a, b]$ the following conditions:

(S1) $f_i(x)$ is continuous, $i = 0, \ldots, n$.

(S2) $f_0(x)$ has constant sign in $[a, b]$.

(S3) $f_i(\bar{x}) = 0 \Rightarrow f_{i-1}(\bar{x}) f_{i+1}(\bar{x}) < 0$, $i = 1, \ldots, n-1$,

$f_n(\bar{x}) = 0 \Rightarrow f_{n-1}(\bar{x}) \neq 0$.

(S4) if $\bar{x}$ is a root of f_n and $h > 0$ small, then

$$\mathrm{sgn}\frac{f_n(\bar{x}-h)}{f_{n-1}(\bar{x}-h)} = -1 \text{ and } \mathrm{sgn}\frac{f_n(\bar{x}+h)}{f_{n-1}(\bar{x}+h)} = +1.$$

(S1) and (S2) are trivial, (S3) can be proved by (5.3) and $f_0 = 1 : f_{i+1}(\bar{x}) = -b_i^2 f_{i-1}(\bar{x})$, so $f_{i-1}(\bar{x}) f_{i+1}(\bar{x}) \leq 0$. If $f_{i-1}(\bar{x}) = 0$, then from (5.3) $f_{i-2}(\bar{x}) = 0 \Rightarrow \cdots \Rightarrow f_0(\bar{x}) = 0$. Contradiction! So $f_{i-1}(\bar{x}) f_{i+1}(\bar{x}) < 0$. For (S4): It is clear for largest root $\bar{x}$, the others follow from induction.

Definition 6.5.1 *A sequence of functions with (S1)-(S4) is called a Sturm chain on* $[a, b]$.

If $x \in [a, b]$, then $f_0(x),\ f_1(x), \cdots, f_n(x)$ are well-defined. Let

$$V(x) = \frac{1}{2}\sum_{i=0}^{n-1} |\mathrm{sgn} f_i(x) - \mathrm{sgn} f_{i+1}(x)|. \tag{5.5}$$

For $f_i(x) \neq 0$, $i = 0, \ldots, n$.$V(x)$ is the number of the sign change of the sequence $f_0(x), \ldots, f_n(x)$. If $f_k(x) = 0$, $1 \leq k \leq n-1$, then $V(x)$ is no differnce, whether sgn 0 is defined by $0, 1$ or -1. Only $\mathrm{sgn} f_n(x)$ must be defined for $f_n(x) = 0$, we set

$$f_n(x) = 0 \ \Rightarrow \mathrm{sgn} f_n(x) := \mathrm{sgn} f_{n-1}(x). \tag{5.6}$$

Theorem 6.5.2 *Let* $f_0, \ldots, f_n$ *be a Sturm chain on* $[a, b]$ *and* $f_n(a) f_n(b) \neq 0$. *Then* $f_n(x)$ *has* $m = V(a) - V(b)$ *roots in* $[a, b]$.

Proof: x runs from a to b, what happens with $V(x)$? $V(x)$ is constant in an interval, if all $f_k(x) \neq 0$, $k = 0, \ldots, n$, $x \in [a, b]$.
(a) x runs through a root $\bar{x}$ of $f_k(x)$, $1 \leq k \leq n-1$. If follows from (S3) that $V(x)$ remains constant.
(b) x runs through a root $\bar{x}$ of $f_n(x)$. Then from (S4) a sign changes is lost. So $V(a) - V(b) =$ the number of roots of $f_k(x)$ in (a, b). ■

For special case as in (5.2), $f_k(\lambda)$ is the characteristic polynomial of A_k. Since $f_k(\lambda) \to \infty$ for $\lambda \to \infty$, so $V(b) = 0$ for large enough b.

Theorem 6.5.3 *If $f_i(x)$ are defined as in (5.2) and $V(x)$ as in (5.5), then holds*

$$V(a) = \textit{the number of eigenvalues of } A \textit{ which are larger than } a.$$

Proof: (1) $f_n(a) \neq 0$. Apply theorem 6.5.2 for large b.
(2) $f_n(a) = 0$, for $\epsilon > 0$ small $\operatorname{sgn} f_i(a+\epsilon) = \operatorname{sgn} f_i(a)$, $i = 0, \cdots, n-1$ and $\operatorname{sgn} f_n(a+\epsilon) = \operatorname{sgn} f_{n-1}(a+\epsilon)$ from (S4). Thus $V(a) = V(a+\epsilon)$ for $\epsilon > 0$. So by theorem 6.5.3 $V(a) =$ the number of eigenvalues of A, which are large than $a + \epsilon$ for arbitrary small $\epsilon > 0$. ■

Calculation of the eigenvalues

Theorem 6.5.3 will be used as the basic tool of the bisection method in locating and separating the roots of $f_n(\lambda)$. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ be the eigenvalues of A as in (5.1) and A is irreducible (i.e., $b_i \neq 0$). Using the Gerschgorin circle theorem 5.2.1 all eigenvalues lie in $[a, b]$, with

$$a = \min_{1 \leq i \leq n} \{a_i - |b_i| - |b_{i-1}|\}$$

$$b = \max_{1 \leq i \leq n} \{a_i + |b_i| + |b_{i-1}|\},$$

where $b_0 = b_n = 0$.

We use the bisection method on $[a, b]$ to divide it into smaller subintervals. Theorem 6.5.3 is used to determine how many roots are contained in a subinterval, and we seek to obtain subintervals that will contain the desired root. If some eigenvalues are nearly equal, then we continue subdividing until the root is found with sufficient accuracy.

Let $a^{(0)}$, $b^{(0)}$ be found with $V(a^{(0)}) \geq k$, $V(b^{(0)}) < k$. Then by theorem 6.5.3 we have $\lambda_k \in (a^{(0)}, b^{(0)}]$.

Determine

$$V\left(\frac{a^{(0)} + b^{(0)}}{2}\right) = v.$$

$$v \geq k \Rightarrow a^{(1)} := \frac{a^{(0)} + b^{(0)}}{2},\ b^{(1)} := b^{(0)}$$

$$v < k \Rightarrow a^{(1)} := a^{(0)},\ b^{(1)} := \frac{a^{(0)} + b^{(0)}}{2},$$

we have $\lambda_k \in (a^{(1)}, b^{(1)}]$. So λ_k is always contained in a smaller interval. The evaluation of $V(\frac{a^{(i)}+b^{(i)}}{2})$ is simply computed by (5.3).

Example 11.1 Consider

$$T = \begin{bmatrix} 2 & 1 & 0 & \dots & 0 \\ 1 & 2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 & 2 \end{bmatrix}.$$

By Gershgorin theorem all eigenvalues lie in $[0, 4]$. 0 and 4 are not the eigenvalues of T (Check!). The roots of T are labeled as

$$0 < \lambda_6 \leq \lambda_5 \leq \ldots \leq \lambda_1 < 4.$$

The roots can be found by continuing the bisection method.

λ	$f_6(\lambda)$	$V(\lambda)$	Comment
0.0	7.0	6	$\lambda_6 > 0$
4.0	7.0	0	$\lambda_1 < 4$
2.0	-1.0	3	$\lambda_4 < 2 < \lambda_3$
1.0	1.0	4	$\lambda_5 < 1 < \lambda_4 < 2$
0.5	-1.421875	5	$0 < \lambda_6 < 0.5 < \lambda_5 < 1$
3.0	1.0	2	$2 < \lambda_3 < 3 < \lambda_2$
3.5	-1.421875	1	$3 < \lambda_2 < 3.5 < \lambda_1 < 4$

Remark 5.1 Although all roots of a tridiagonal matrix may be found by this technique, it is generally faster in that case to use the QR algorithm. With large matrices, we usually do not want all roots, so the method of this section are preferable. If we only want some certain specific roots, for example, the five largest or all roots in a given interval, it is easy to locate them by using theorem 6.5.3.

6.5.2 Rayleigh Quotient Iteration

Suppose $A \in \mathbf{R}^{n\times n}$ is symmetric and $x \neq 0$ is a given vector. A simple differentiation reveals that

$$\lambda = R[x] \equiv \frac{x^T A x}{x^T x} \tag{5.7}$$

minimizes $\|(A - \lambda I)x\|_2$. The scalar $r(x)$ is called the Rayleigh quotient of x. If x is an approximate eigenvector, then $r(x)$ is a reasonable choice for the corresponding eigenvalue. On the other hand, if λ is an approximate eigenvalue, then inverse iteration tells us that the solution to $(A - \lambda I)x = b$ will almost always be a good approximate eigenvector.

Combining these two ideas lead to the Rayleigh-quotient iteration:

$$\begin{array}{l} \text{Given } x_0 \text{ with } \|x_0\|_2 = 1. \\ \text{For } k = 0, 1, \ldots \\ \mu_k = R[x_k] \\ \text{ Solve } (A - \mu_k I)z_{k+1} = x_k \text{ for } z_{k+1} \\ x_{k+1} = z_{k+1}/\|z_{k+1}\|_2. \end{array} \tag{5.8}$$

Parlett (1974) has shown that (5.8) converges globally and the loccally cubically. (See also Chapter I).

6.5.3 Orthogonal Iteration with Ritz Acceleration

$$\begin{array}{l}\text{Given } Q_0 \in \mathbf{R}^{n\times p} \text{ with } Q_0^T Q_0 = I_p.\\ \text{For } k = 0, 1, \ldots\\ Z_k = AQ_{k-1},\\ Q_k R_k = Z_k \text{ (QR-decomposition).}\end{array} \tag{5.9}$$

Let $Q^T AQ = diag(\lambda_i)$ be the Schur decomposition of A and $Q = [q_1, \ldots, q_n]$, and $|\lambda_1| > |\lambda_2| > \ldots |\lambda_n|$. If follows from theorem 5.3.4 that if

$$d = dist[D_\rho(A), R(Q_0)] < 1,$$

then

$$dist[D_\rho(A), R(Q_k)] \leq \frac{1}{\sqrt{1-d^2}} |\frac{\lambda_{p+1}}{\lambda_p}|^k.$$

We know that (Stewart 1976) if $R_k = [r_{ij}^{(k)}]$ then

$$|r_{ii}^{(k)} - \lambda_i| = O(|\frac{\lambda_{i+1}}{\lambda_i}|^k), \quad i = 1, \ldots, p.$$

This can be an unacceptably slow rate of convergence if λ_i and λ_{i+1} are of nearly equal modulus. This difficulty can be surmounted by replacing Q_k with its Ritz Vectors at each step:

$$\begin{array}{l}\text{Given } Q_0 \in \mathbf{R}^{n\times p} \text{ with } Q_0^T Q_0 = I_p.\\ \text{For } k = 0, 1, \ldots\\ Z_k = AQ_{k-1},\\ \tilde{Q}_k R_k = Z_k \ (QR \text{ decomposition}),\\ S_k = \tilde{Q}_k{}^T A \tilde{Q}_k,\\ U_k^T S_k U_k = D_k \text{ (Schur decomposition)},\\ Q_k = \tilde{Q}_k U_k.\end{array} \tag{5.10}$$

It can be shown that if

$$D_k = diag(\theta^{(k)}, \ldots, \theta_p^{(k)}) \quad \text{and} \quad |\theta_1^{(k)}| \geq \cdots \geq |\theta_p^{(k)}|,$$

then

$$|\theta_i^{(k)} - \lambda_i(A)| = |\frac{\lambda_{p+1}}{\lambda_i}|^k, \quad i = 1, \ldots, p.$$

Thus the Ritz values $\theta_i^{(k)}$ converge in a more favorable rate than the $r_{ii}^{(k)}$ in (5.9). For details, see Stewart (1969) and Parlett's book chapters 11 and 14.

6.6 Generalized Definite Eigenvalue Problem $Ax = \lambda Bx$

6.6.1 Generalized definite eigenvalue problem

$$Ax = \lambda Bx, \tag{6.1}$$

where $A, B \in \mathbf{R}^{n\times n}$ are symmetric and B is positive definite. (In practice A, B are very large and sparse).

Theorem 6.6.1 *The eigenvalue problem (6.1) has n real eigenvalues λ_i associated with egienvectors x_i satisfying*

$$Ax_i = \lambda_i Bx_i,\ i = 1, \ldots, n. \tag{6.2}$$

Here $\{x_i\}_{i=1}^n$ can be chosen such that $x_i^T Bx_j = \delta_{ij}$ (B-orthogonal), $i, j = 1, \ldots, n$.

Proof: Let $B = LL^T$ be the Cholesky decomposition of B. Then $Ax_i = \lambda_i Bx_i \Longleftrightarrow Ax_i = \lambda_i LL^T x_i \Longleftrightarrow L^{-1}AL^{-T}(L^T x_i) = \lambda_i (L^T x_i) \Longleftrightarrow Cz_i = \lambda_i z_i$, where $C \equiv L^{-1}AL^{-1}$ symmetric and $z_i \equiv L^T x_i$. Since λ_i are the eigenvalues of the symmetric matrix C, they are real. The vectors z_i can be chosen pairwisely orthogonal, i.e., $z_i^T z_i = \delta_{ij} = x_i^T LL^T x_j = x_i^T Bx_j$. ■

Let $X = [x_1, \cdots, x_n]$. Then from above we have $X^T BX = I$ and $(X^T AX)_{ij} = x_i^T Ax_j = \lambda_j x_i^T Bx_j = \lambda_j \delta_{ij}$ which implies $X^T AX = \Lambda = diag(\lambda_1, \cdots, \lambda_n)$. That is, A, B are simultaneously diagonalizable by a congruence transformations.

Numerical methods for (6.1):
(a) Bisection method,
(b) Coordinate relaxation,
(c) Method of steepest descent.

(a) Bisection methods:

Basic tool: Sylvester law of inertia

Definition 6.6.1 *Two real, symmetric matrices A, B are called congruent, if there exists a nonsingular C such that*

$$A = C^T BC. \tag{6.3}$$

We denote it by $A \overset{c}{\sim} B$.

Defintion 6.6.2 *The inertia of a symmetric matrix A is a triplet of integers*

$$in(A) = (\pi(A), \nu(A), \delta(A)) \tag{6.4}$$

$\pi(A) =$ *the number of positive eigenvalues of A (geometry multiplicity),*
$\nu(A) =$ *the number of negative eigenvalues of A (geometry multiplicity),*
$\delta(A) = n - rank(A) =$ *the number of zero eigenvalues of A.*

Theorem 6.6.2 (Sylvester law of inertia) *Two real, symmetric matrices are congruent if and only if they have the same inertia.*

Proof: (1) A, B real and symmetric. Suppose $in(A) = in(B)$, there exist orthogonal U and V such that $UAU^T = \Lambda_1 = diag(\lambda_1(A), \cdots, \lambda_n(A))$ with $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ and $VBV^T = \Lambda_2 = diag(\lambda_1(B), \cdots, \lambda_n(B))$ with $\lambda_1(B) \geq \cdots \geq \lambda_n(B)$.

Claim: Λ_1 is congruent to Λ_2. Since $in(A) = in(B)$, it holds either $\lambda_i(A)\lambda_i(B) > 0$ or $\lambda_i(A) = \lambda_i(B) = 0$. Set $D = diag(d_i)$, where

$$d_i = \begin{cases} \sqrt{\frac{\lambda_i(A)}{\lambda_i(B)}}, & \text{if } \lambda_i(A)\lambda_i(B) > 0 \\ 1, & \text{if } \lambda_i(A)\lambda_i(B) = 0 \end{cases}.$$

Then $D^T\Lambda_2 D = \Lambda_1$, so $A \overset{c}{\sim} B$.

(2) Suppose $A \overset{c}{\sim} B$. Claim: $in(A) = in(B)$. Let $A = C^TBC$, $UAU^T = \Lambda_1$ and $VBV^T = \Lambda_2$ as above. These imply $\Lambda_1 = P^T\Lambda_2 P$, where $P = V^TCU^T$. Assume that $in(A) \neq in(B)$. Clearly $\delta(A) = \delta(B) \Rightarrow \pi(A) \neq \pi(B)$. Without loss of generality we can suppose $\pi(A) < \pi(B)$. The homogenous linear system

$$\begin{cases} x_i = 0, & i = 1, \cdots, \pi(A), \\ (Px)_i = 0, & i = \pi(B) + 1, \cdots, n, \end{cases} \tag{6.5}$$

has a nonzero solution $x \neq 0$, since it has fewer than n equations. With this x we have

$$\begin{aligned} 0 &\geq \sum_{i=1}^{n} \lambda_i(A)x_i^2 = x^T\Lambda_1 x = x^TP^T\Lambda_2 Px \\ &= \sum_{i=1}^{n} \lambda_i(B)(Px)_i^2 > 0 \\ &= \sum_{i=1}^{\pi(B)} \lambda_i(B)(Px)_i^2. \end{aligned}$$

That is, there is an i $(1 \leq i \leq \pi(B))$ with $(Px)_i \neq 0$, contradiction!

Second Part of Proof:

Show that B and C^TBC have the same inertia. Because they have the same rank, it is sufficient to show that: $\pi(B) = \pi(C^TBC)$.

If $\lambda_r(B) > 0$, let $Bq_i = \lambda_i(B)q_i$ and $S_0 = span\{C^{-1}q_q, \cdots, C^{-1}q_r\}$. Then

$$\begin{aligned} \lambda_r(C^TBC) &= \max_{dim\ S=r} \min_{x\in S,\ x\neq 0} \frac{x^TC^TCx}{x^Tx} \geq \min_{x\in S_0,\ x\neq 0} \frac{x^TC^TBCx}{x^Tx} \\ &= \min_{x\in S,\ x\neq 0} \frac{x^TC^TBCx}{x^TC^TCx} \frac{x^TC^TCx}{x^Tx} \\ &\geq \min_{x\in S_0,\ x\neq 0} \frac{x^TC^TBCx}{x^TC^TCx} \min_{x\in \mathbf{R}^{n\times n},\ x\neq 0} \frac{x^TC^TCx}{x^Tx} \\ &= \lambda_r(B)\sigma_n(C)^2 > 0, \quad (\text{Since } x \in S_0 \Rightarrow Cx \in Span(q_1, \cdots, q_r)) \end{aligned}$$

where $\sigma_1(C) \geq \cdots \geq \sigma_n(C) > 0$ are the singular values of C. So we have $\lambda_r(C^TBC) > 0$ and $\pi(C^TBC) \geq \pi(B)$. Exchange the role of B and C^TBC we obtain $\pi(C^TBC) = \pi(B)$.

Important inequality:

From above we have $\lambda_r(C^TBC) \geq \lambda_r(B)\sigma_n^2(C)$. Change B and C^TBC we then obtain

$$\lambda_r(B) \geq \lambda_r(C^TBC)\sigma_n^2(C^{-1}) = \lambda_r(C^TBC)\frac{1}{\sigma_1^2(C)}.$$

This imply

$$\sigma_1^2(C) \geq \frac{\lambda_r(C^TBC)}{\lambda_r(B)} \geq \sigma_n^2(C). \tag{6.6}$$

It holds also for the negative eigenvalues of B and C^TBC. ■

Corollary 6.6.3 *If $A = C^TBC$, C nonsingular ($A \overset{c}{\sim} B$), then it holds for nonzero eigenvalues*

$$\sigma_1^2(C) \geq \frac{\lambda_r(A)}{\lambda_r(B)} \geq \sigma_n^2(C).$$

Lemma 6.6.4 *A nonsigular, real and symmetric and has a LR-decomposition*

$$A = LR, \tag{6.7}$$

where L is lower triangular with $l_{ii} = 1$ and R is upper triangular with $r_{ii} \neq 0$, $i = 1, \cdots, n$. Then holds

$$\pi(A) = \#\{i : r_{ii} > 0\} \quad \text{and} \quad \nu(A) = \#\{i : r_{ii} < 0\}.$$

Proof: Let $D = diag(r_{ii})$. Then $\tilde{R} = D^{-1}R$ has "one" on the diagonal. This implies

$$A = LR = LD\tilde{R} = A^T = \tilde{R}^TDL^T.$$

The decomposition $A = \tilde{L}\tilde{D}\tilde{R}$, where $\tilde{L}, \tilde{R}$ has "one" on the diagonal is unique, therefore $L = \tilde{R}^T$. So

$$A = LDL^T \Longrightarrow A \overset{c}{\sim} D \Longrightarrow in(A) = in(D).$$

But $\pi(D) = \#\{i : r_{ii} > 0\}$, the assertion is proved. ■

Theorem 6.6.5 *Let A, B be real, symmetric and B positive definite, α be a given real number. Then holds*

$$\begin{aligned}
\pi(A - \alpha B) &= \#\{\textit{eigenvalues of (6.1) larger than } \alpha\} \\
\nu(A - \alpha B) &= \#\{\textit{eigenvalues of (6.1) smaller than } \alpha\} \\
\delta(A - \alpha B) &= \#\{\textit{multiplicity of } \alpha \textit{ as an eigenvalues of (6.1)}\}
\end{aligned}$$

Proof: $Ax = \lambda Bx \Longleftrightarrow Cy = \lambda y$, where $C = L^{-1}AL^{-1}$, $B = LL^T$ and $L^Tx = y$. By theorem 6.6.2 (Sylvester law of inertia) we have

$$in\{(A - \alpha B)\} = in\{L^{-1}(A - \alpha B)L^{-1}\} = in\{(C - \alpha I)\}.$$

Since $C - \alpha I$ has the eigenvalues $\lambda_i - \alpha > 0$, we have

$$\pi(A - \alpha B) = \#\{i : \lambda_i - \alpha > 0\} = \#\{i : \lambda_i > \alpha\}.$$

Similarly, we also have the assertions for $\nu(A - \alpha B)$ and $\delta(A - \alpha B)$. ■

Remark 6.1 Theorem 6.6.5 leads to a bisection method for (6.1). If $[a, b]$ is an interval, which contains the desired eigenvalues, then by calculation of $in(A - \frac{a+b}{2}B)$ we know that the desired eigenvalues lie in $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}, b]$. It requires the LU decomposition of $A - \alpha B$, which in general is indefinite.

(b) Methods of Coordinate relaxatoin:

This method requires only the calculation of Ax and Bx. Consider the generalized Rayleigh quotient

$$R[x] = \frac{x^TAx}{x^TBx}. \tag{6.8}$$

Let $z = L^Tx$, $C = L^{-1}AL^{-T}$ and $B = LL^T$ $(Ax = \lambda Bx)$. C is symmetric, let $Cu_i = \lambda_i u_i$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. By theorem 6.1.6 we have

$$\lambda_i = \max\{R[z, C] = \frac{z^TCz}{z^Tz} : z \perp u_i,\ j < i,\ z \neq 0\}.$$

From (6.8) follows that

$$R[x] = \frac{x^TAx}{x^TBx} = \frac{z^TL^{-1}AL^{-T}z}{z^Tz} = \frac{z^TCz}{z^Tz}.$$

Therefore we have the following new version of connection between the eigenvalues and Rayleigh quotient of generalized definite eigenvalues problem (6.1).

Theorem 6.6.6 *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $Ax = \lambda Bx$ satisfying $Ax_i = \lambda_i Bx_i, i = 1, \cdots, n$. It holds*

$$\lambda_i = \max\{R[x] = \frac{x^TAx}{x^TBx} : \quad x^TBx_j = 0, \quad j < i,\ x \neq 0\}. \tag{6.9}$$

Proof: $Ax_i = \lambda_i Bx_i \Longleftrightarrow Cu_i = \lambda_i u_i, u_i = L^Tx_i$. Let $z = L^Tx$, then $z \perp u_j \Longleftrightarrow z^Tu_j = 0 \Longleftrightarrow x^TLL^Tx_j = 0 \Longleftrightarrow x^TBx_j = 0$. These imply that

$$\begin{aligned} & \{\frac{z^TCz}{z^Tz} : z \perp u_j, j < i, z \neq 0\} \\ = \ & \{\frac{x^TAx}{x^TBx} : x^TBx_j = 0, j < i, x \neq 0\}. \end{aligned}$$

Take maximum and from (1.5) follows (6.9). ■

Similarly, theorem 6.1.6 (Courant-Fischer) can be transfered to:

Theorem 6.6.7 *For the eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$ of $Ax = \lambda Bx$ it holds*

$$\lambda_i = \min_{\substack{\{p_1,\cdots,p_{i-1}\},\\ 1\leq j<i, p_j\neq 0}} \max_{\substack{p_j^T x=0,\\ 1\leq j<i, x\neq 0}} \frac{x^TAx}{x^TBx}, \tag{6.10}$$

$$\lambda_i = \min_{dimS=n+1-i} \max_{x\in S, x\neq 0} \frac{x^TAx}{x^TBx}, \tag{6.11}$$

$$\lambda_i = \max_{dimS=i} \min_{x\in S, x\neq 0} \frac{x^TAx}{x^TBx}. \tag{6.12}$$

■

Theorem 6.1.7 (Separation theorem) can be transfered to:

Theorem 6.6.8 *A, B are real, symmetric and B is positive definite. A_{n-1} and B_{n-1} are obtained by scratching the last row and column of A and B respectively. For the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ of $Ax = \lambda Bx$ and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-1}$ of $A_{n-1}x = \tilde{\lambda} B_{n-1}x$ it holds*

$$\lambda_{s+1} \leq \mu_s \leq \lambda_s, \quad s = 1, \cdots, n-1 \tag{6.13}$$

and

$$\lambda_1 = \max_{x\neq 0} R[x], \quad \lambda_n = \min_{x\neq 0} R[x]. \tag{6.14}$$

■

Problem: How to compute the smallest eigenvalue λ_n and its associated eigenvector?
Ideal: Minimize Rayleigh quotient $R[x]$ on a two dimension subspace.
Basic Problem: Given two linearly independent vectors x, y. Minimize $R[x]$ on the from x and y generated subspace generated by x and y.

Let $x' = \phi x + \gamma y$, then

$$R[x'] = \frac{(\phi x+\gamma y)^T A(\phi x+\gamma y)}{(\phi x+\gamma y)^T B(\phi x+\gamma y)} = \frac{\phi^2\alpha + 2\phi\gamma f + \gamma^2 p}{\phi^2\beta + 2\phi\gamma g + \gamma^2 q}, \tag{6.15}$$

where $\alpha = x^TAx,\ \beta = x^TBx,\ f = x^TAy,\ g = x^TBy,\ p = y^TAy$ and $q = y^TBy$. Let

$$\tilde{A} = \begin{bmatrix} \alpha & f \\ f & p \end{bmatrix}, \tilde{B} = \begin{bmatrix} \beta & g \\ g & q \end{bmatrix}, \tilde{x} = \begin{bmatrix} \phi \\ \gamma \end{bmatrix}. \tag{6.16}$$

Then

$$R[x'] = \frac{\tilde{x}^T\tilde{A}\tilde{x}}{\tilde{x}^T\tilde{B}\tilde{x}},$$

where $\tilde{A}, \tilde{B}$ are symmetric and B is positive definite. Applying (6.14) to $\tilde{A}$ and $\tilde{B}$ we get that $R[x']$ has the minimum R', where R' is the smallest eigenvalue of the problem $\tilde{A}\tilde{x} = \tilde{\lambda}\tilde{B}\tilde{x}$. That is,

$$det(\tilde{A} - R'\tilde{B}) = 0, \text{ quadratic equation in } R'. \tag{6.17}$$

Compute the associated eigenvector (to R') $\tilde{x} = (\phi, \gamma)^T$ from one of the following equations:

$$(\alpha - R'\beta)\phi + (f - R'g)\gamma = 0 \tag{6.18}$$
$$(f - R'g)\phi + (p - R'q)\gamma = 0 \tag{6.19}$$

$$\Longleftrightarrow (\tilde{A} - R'\tilde{B}) \begin{bmatrix} \phi \\ \gamma \end{bmatrix} = 0.$$

Case 1: $p - R'q \neq 0$ ($p/q = y^T Ay/y^T By = R[y] > R'$):

From (6.19) implies $\phi \neq 0$. Set $\phi = 1$. From (6.19) follows

$$\gamma = -\frac{f - R'g}{p - R'q} \tag{6.20}$$

and that

$$x' = x + \gamma y \tag{6.21}$$

is the solution of the basic problem. Case 1 is called normal case.

Case 2: $p - R'q = 0$: This implies $f - R'g = 0$, because

$$0 = det(\tilde{A} - R'\tilde{B}) = (\alpha - R'\beta)(p - R'q) - (f - R'g)^2.$$

(a) If $\alpha - R'\beta \neq 0$, then $\phi = 0$ and γ is arbitray. Set $x' = y$.
(b) If $\alpha - R'\beta = 0$, then $\tilde{A} = R'\tilde{B} \Longrightarrow R[\tilde{x}] = R'$ for all $\tilde{x} \in Span(x, y)$. Set $x' = x$.

The method of coordinate relaxation

> Given a starting vecor $y_1 \neq 0$.
> y_{i+1} is determined by y_i as follows:
> Set $x = y_i, y = e_k, k = i \ mod \ n$ and
> Solve the basic problem with respect to x and y.
> Let x' be the solution. Set $y_{i+1} = x'/|x'|$.

We obtain the sequence of vectors $y_1, y_2, y_3, \cdots$ such that

$$R[y_1] \geq R[y_2] \geq R[y_3] \geq \cdots \geq \lambda_n.$$

Remark to the computational cost

(1) Compute $\tilde{A}, \tilde{B}$: compute

$$p = y^T Ay = e_k^T Ae_k = a_{kk}, \quad q = e_k^T Be_k = b_{kk},$$
$$u = Ax \text{ and } v = Bx, \text{ and then}$$
$$f = y^T Ax = e_k^T u = u_k, \quad g = y^T Bx = e_k^T v = v_k,$$
$$\alpha = x^T Ax = x^T u \text{ and } \beta = x^T v.$$
Construct $\tilde{A}$ and $\tilde{B}$.

(2) Solve the quadratic equation $det(\tilde{A} - R'\tilde{B}) = 0$ in R'.

(3) Solve $x' = x + \gamma e_k$.

(4) Ax' and Bx' (for the next step) can be computed implicitly. We use the following updating:

$$Ax' = Ax + \gamma Ae_k, \quad (Ax')_j = u_j + \gamma a_{jk},$$
$$Bx' = Bx + \gamma Be_k, \quad (Bx')_j = v_j + \gamma b_{jk}.$$

Remark 6.2 If $R[y_1] < \min_i(a_{ii}/b_{ii}) = \min_i R[e_i]$, then it happens only normal case:

$$R'q - p = R'b_{kk} - a_{kk} \leq R[y_1]b_{kk} - a_{kk} < 0.$$

Since $R[y_1] < a_{kk}/b_{kk}$, so $R'q - p \neq 0$, a normal case!

Theorem 6.6.9 *Let*

$$R[y_1] < \min_i \frac{a_{ii}}{b_{ii}}, \tag{6.22}$$

Then it holds

$$\lim_{i\to\infty} R[y_i] = \lambda. \tag{6.23}$$

Here λ is an eigenvalue of (6.1) $Ax = \lambda Bx$, and each accumulation point of $\{y_i\}$ is the associated eigenvector to λ.

Corollary 6.6.10 *If (6.22) holds and*

$$R[y_1] < \lambda_{n-1}, \tag{6.24}$$

Then $\lim_{i\to\infty} R[y_i] = \lambda_n$. If λ_n is simple, then holds:

$$\lim_{i\to\infty} y_i = y \text{ exists and } y \text{ is the eigenvector to } \lambda_n.$$

Proof of theorem 6.6.9 Only normal case!

y_{i+1} is a function of $e_k(k = i \ mod \ n)$ and y_i. Let $y_{i+1} = T_k(y_i)$. The function T_k is continuous in y_i, since for fix y the solution x' of basic problem depends continously on the given x. (normal case!)

For $R[y_1] \geq R[y_2] \geq \cdots \geq \lambda_n$ there exists the limit point λ. Show that:

$$\lambda = \lim_{i\to\infty} R[y_i] \text{ is an eigenvalue.}$$

In addition we show that an accumulation point y of the sequence $\{y_i\}$ satisfies $Ay = \lambda By$.

Let $y_{r(i)}$ be the convergence subsequence of $\{y_i\}$, i.e., $\lim_{i\to\infty} y_{r(i)} = y$. Without loss of generality there are infinite $r(i)$ satisfying $1 = r(i) \ mod \ n$. So

$$y = \lim_{i\to\infty} y_{nk(i)+1} \text{ and } R[y] = \lambda.$$

Since T_1 is continuous, where

$$T_1 y = \lim_{i\to\infty} T_1 y_{nk(i)+1} = \lim_{i\to\infty} y_{nk(i)+2}$$

which implies $R[T_1y] = \lambda$. Thus, $R[T_1y] = R[y] = \lambda \Longrightarrow \gamma = 0 \Longrightarrow y = T_1y$ and $f - \lambda g = 0$. So $f = (Ay)_1$ and $g = (By)_1 \Longrightarrow (Ay)_1 = \lambda(By)_1$. Also T_2 is continuous, we have

$$T_2T_1y = T_2y = \lim_{i\to\infty} T_2y_{nk(i)+2} = \lim_{i\to\infty} y_{nk(i)+3},$$

then $\lambda = R[y] = R[T_2y]$. As above we also have $\gamma = 0$ and then $y = T_2y$. So $f = \lambda g$, thus $(Ay)_2 = \lambda(By)_2$, and so on. It follows $Ay = \lambda By$. ■

Proof of Corollaray 6.6.10
The first part is trivial, since λ_n is the unique eigenvalue smaller than λ_{n-1}. The normalized eigenvectors to λ_n are $\pm x/|x|$, where $Ax = \lambda_n x$. Two possible accumulation points are separate. Let $y_i \approx x/|x|$, then $Ay_i \approx \lambda_n By_i$. This follows $f \approx \lambda_n g$, so $\gamma \approx 0$, thus $y_{i+1} \approx y_i$. A second accumulation point can not appear. ■

As relaxation method by solving linear system, we introduce an "overcorrect" $x' = x + \omega\gamma y$ $(1 < \omega < 2)$ for the csae 1 instead of $x' = x + \gamma y$. We describe the above discussion as the following algorithm:

Algorithm 6.1 (Coordinate over relaxation method to determine the smallest eigenvalue of symmetric generalized definite problem $Ax = \lambda Bx$)

Let $A, B \in \mathbf{R}^{n\times n}$ be symmetric and B positive definite.
Step 1: Choose a relaxation factor $\omega \in (1, 2)$, tolerance $\delta, \epsilon \in \mathbf{R}_+$ and a starting vector $x \in \mathbf{R}^n\backslash\{0\}$. Compute $a := x^TAx, b := x^TBx$ and $r := a/b$.

Step 2: Set $R_{max} := R_{min} := r$.

For $j = 1, 2, \cdots, n$
Compute $f : \sum_{k=1}^n a_{jk}x_k, \quad g := \sum_{k=1}^n b_{jk}x_k, \quad p := a_{jj}, \quad q := b_{jj}$
Determine the smallest eigenvalue r_1 of

$$\left(\begin{bmatrix} a & f \\ f & p \end{bmatrix} - \lambda \begin{bmatrix} b & g \\ g & q \end{bmatrix}\right) \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = 0.$$

(2.1) If $|p - r_1q| > \epsilon$, then set

$$\beta := -\omega\tfrac{f-r_1g}{g-r_1q}, x_j := x_j + \beta, a := a + 2\beta f + \beta^2 p,$$
$$b := b + 2\beta g + \beta^2 g, r := \tfrac{a}{b};$$

(2.2) If $|p - r_1q| \le \epsilon$, and $|a - r_1b| > \epsilon$ then set

$$x := e_j, a := p, b := q, r := \frac{a}{b}$$

(2.3) If $|p - r_1q| \le \epsilon$, and $|a - r_1b| \le \epsilon$ then stop. Set

$$R_{max} := \max(r, R_{max}) \text{ and } R_{min} := \min(r, R_{min}).$$

Step 3: If $\frac{R_{min}}{R_{max}} \leq 1-\delta$, then go to step 2, otherwise stop.

A detail discussions for determining optimal ω can be found in:
H.R.Schwarz: Numer. Math. 23, 135-151 (1974).
H.R.Schwarz: Finite Elemente, Teubner Verlag.

(c) Methods of steepest descent

Recall that at a point x_k the function $\phi : \mathbf{R^n} \longrightarrow \mathbf{R}$ decreases most rapidly in the direction of the negative gradient $-\bigtriangledown\phi(x_k)$. The method is called the gradient or steepest descent method. Here we have

$$\phi(x) = R[x] = \frac{x^T Ax}{x^T Bx}.$$

It holds

$$\text{Grad}\phi(x) = \frac{2[(x^T Bx)Ax - (x^T Ax)Bx]}{(x^T Bx)^2} = \frac{2}{x^T Bx}(Ax - R[x]Bx). \tag{6.25}$$

Thus, Grad $R[x]$ (=Grad $\phi(x)$) $= 0 \Longleftrightarrow R[x]$ is the eigenvalue and x is the associated eigenvector $\overset{def}{\Longleftrightarrow} x$ is stationary point of $R[x]$.

Methods of steepest descent:

Given $y_1 \neq 0$. y_{i+1} is determined by y_i.
(1) Search direction

$$p_i = (A - R[y_i]B)y_i. \tag{6.26}$$

If $p_i = 0$ stop, otherwise
(2) Solve the basic problem with $x = y_i$ and $y = p_i$.
If x' is the solution, then set

$$y_{i+1} = \frac{x'}{\|x'\|}, \tag{6.27}$$

Go to (1).

Lemma 6.6.11 *Let $B = I$. Then holds*

$$p_i^T(A - R'B)y_i = p_i^T p_i, R' = R[y_{i+1}], \tag{6.28}$$

$$p_i^T(A - R'B)p_i > 0, \ \textit{if } p_i \neq 0. \tag{6.29}$$

Especially, it happens only normal case, thus the function $T(y_i) = y_{i+1} = T(y_i)$ is continuous.

Proof: Since $p_i^T y_i = 0$ (by computation!), we have

$$\begin{aligned} p_i^T(A - R'B)y_i &= p_i^T(A - R[y_i]B)y_i + (R[y_i] - R')p_i^T By_i \\ &= p_i^T p_i. \end{aligned}$$

If $p_i \neq 0$, then $f - R'g = p_i^T Ay_i - R'p_i^T By_i = p_i^T p_i > 0$ (by (6.28)). From (6.18) and $f - R'g \neq 0 \Rightarrow \phi \neq 0$. Hence the minimum is not at $y = p_i$, so $R[p_i] > R'$. Thus

$$p_i^T(A - R'B)y_i = p_i^T(A - R[p_i]B)p_i + (R[p_i] - R')p_i^T Bp_i > 0.$$

So (6.29) holds. i.e. $p - R'q \neq 0 \Rightarrow$ normal case! $\Rightarrow T$ is continuous. ■

Theorem 6.6.12 *Let $B = I$, and the sequence of vectors $\{y_i\}$ is obtained by the method of steepest descent (6.26),(6.27). Then it holds with $r_i = R[y_i]$ that*

(1) $\lim_{i\to\infty} r_i = \lambda$ *is an eigenvalue of A .*

(2) Each accumulation point of $\{y_i\}$ is the eigenvector of A corresponding to λ.

(3) If $y_1 = \sum_{k=1}^{n} \alpha_k x_k$ is the expansion of the starting vector by normalized eigenvectors $\{x_k\}_{k=1}^{n}$ of A, ($Ax_i = \lambda_i x_i$ with $\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$) and $\alpha_n \neq 0$, then it holds

$$\lim_{i\to\infty} r_i = \lambda_n.$$

Proof: Since $r_1 \geq r_2 \geq \cdots \geq \lambda_n$, there exists the limit point λ with $lim_{i\to\infty} r_i = \lambda$. Let z be an accumulation point of $\{y_i\}_{i\in N}$, i.e.,

$$z = \lim_{i\to\infty} y_{n(i)}, \quad R[z] = \lim_{i\to\infty} R[y_{n(i)}] = \lambda.$$

Since $T(T : y_i \to y_{i+1})$ is continuous, so

$$\lim_{i\to\infty} Ty_{n(i)} = \lim_{i\to\infty} y_{n(i)+1} = T \lim_{i\to\infty} y_{n(i)} = Tz.$$

This implies

$$R[Tz] = \lim_{i\to\infty} R[y_{n(i)+1}] = \lambda.$$

From $R[Tz] = R[z]$ and $\gamma = 0$ we have $Tz = z$. (Since Grad $R[z] = 0, z$ is the eigenvector to $R[z] = \lambda$). Thus (1),(2) are established.

Claim (3): Let $y_i = \sum_{k=1}^{n} \alpha_k^i x_k$. Prove that α_k^{i+1} is determined by α_k^i. Since

$$p_i = (A - r_i I)y_i = \sum_{k=1}^{n} (\lambda_k - r_i)\alpha_k^i x_k,$$

from (6.28)(6.29) follows

$$-\gamma_i = \frac{f - r_{i+1}g}{p - r_{i+1}q} = \frac{p_i^T(A - r_{i+1}B)y_i}{p_i^T(A - r_{i+1}B)p_i} > 0.$$

Since

$$y_i + \gamma_i p_i = \sum_{k=1}^{n} \alpha_k^i [1 + \gamma_i(\lambda_k - r_i)]Bx_k,$$

we get

$$y_{i+1} = \frac{y_i + \gamma_i p_i}{\|y_i + \gamma_i p_i\|} = \sum_{k=1}^{n} \frac{\alpha_i^k (1 + \gamma_i(\lambda_k - r_i))}{(\sum_{s=1}^{n} (\alpha_s^i)^2 (1 + \gamma_i(\lambda_k - r_i))^2)^{1/2}} Bx_k.$$

This implies

$$\alpha_r^{i+1} = \frac{\alpha_r^i(1+\gamma_i(\lambda_r - r_i))}{\sqrt{\sum_{s=1}^n (\alpha_s^i)^2(1+\gamma_i(\lambda_s - r_i))^2}}. \tag{6.30}$$

We then have

$$\frac{\alpha_n^{i+1}}{\alpha_r^{i+1}} = \frac{\alpha_n^i}{\alpha_r^i}\left(\frac{1+\gamma_i(\lambda_n - r_i)}{1+\gamma_i(\lambda_r - r_i)}\right). \tag{6.31}$$

Assume that $\{r_i\}$ does not converge to λ_n, but to $\lambda_r > \lambda_n$. Then

$$y_{n(i)} \to \pm x_r \Longrightarrow \alpha_r^{n(i)} \to \pm 1 \quad \text{and} \quad \alpha_n^{n(i)} \to 0.$$

On the other hand, since

$$\lambda_r - r_i < 0, \quad \lambda_n - r_i < 0, \quad \gamma_i < 0 \text{ and } \lambda_n - r_i < \lambda_r - r_i,$$

we have

$$\frac{1+\gamma_i(\lambda_n - r_i)}{1+\gamma_i(\lambda_r - r_i)} > 1.$$

From (6.31) follows that

$$\left|\frac{\alpha_n^{i+1}}{\alpha_r^{i+1}}\right| > \left|\frac{\alpha_n^i}{\alpha_r^i}\right|.$$

This contradicts that $\alpha_n^{n(i)} \to 0$. ■

Further methods for the symmetric eigenvalue problem $Ax = \lambda Bx$

$A, B \in \mathbf{R}^{n\times n}$ are symmetric and B is positive definite. Reduction to the ordinary eigenvalue problem:

(1) $B^{-1}Ax = \lambda x$, the symmetry is lost.

(2) $B = LL^T, L^{-1}AL^{-T}z = \lambda z$ with $L^Tx = z$, Cholesky method.

Remark: For a given vector z we can compute $L^{-1}AL^{-T}z$ as follow: Comupte z_1 from $L^Tz_1 = z$ by backward substitution. Then $z_2 = Az_1$. Compute z_3 from $Lz_3 = z_2$ by forward substitution. We can use the sparsity of B (also L) and A.

Caution: $L^{-1}AL^{-T}$ is in general dense.

(3) Theoretically, B has (unique) positive definite square root $B^{1/2}$, i.e., $B^{1/2}B^{1/2} = B$, $B^{-1/2}AB^{-1/2}z = \lambda z$. Computation of $B^{1/2}$ is expensive. Let $B = UDU^T$, where U is orthogonal and D is diagonal with $D > 0$. Then

$$B = (UD^{1/2}U^T)(UD^{1/2}U^T) \Rightarrow B^{1/2} = UD^{1/2}U^T,$$

where $D = diag(d_i)$ and $D^{1/2} = diag(\sqrt{d_i})$.

Consider

$$Ax = \lambda Bx, A, B \text{ symmetric and } B \text{ positive definite.}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ be the eigenvalues of (6.1). Recall that the power method and the inverse iteration for $B = I$:
Power method:

$$\begin{array}{l} \text{Given } x_0 \neq 0, \\ \text{for } i = 0, 1, 2, \cdots, \\ \quad y_{i+1} = Ax_i \ , \quad k_{i+1} = \|y_{i+1}\|, \\ \quad x_{i+1} = y_{i+1}/k_{i+1}. \end{array} \tag{6.32}$$

x_i converges to the eigenvector of the eigenvalue λ_1 and $k_i \to |\lambda_1|$ as $i \to \infty$.

Inverse power method:

$$\begin{array}{l} \text{Given } x_0 \neq 0, \\ \text{for } i = 0, 1, 2, \cdots, \\ \quad \sigma_i = x_i^TAx_i/x_i^Tx_i \text{ Rayleigh quotient} \\ \quad (A - \sigma_i I)x_{i+1} = k_{i+1}x_i, \\ \quad k_{k+1} \text{ is chosen so that } \|x_{i+1}\| = 1. \end{array} \tag{6.33}$$

Cubic convergence.

Transfer to the problem (6.1):

Power method (6.32) $\Leftrightarrow A \leftrightarrow B^{-1}A$:

$$\begin{array}{l} \text{Given } x_0 \neq 0, \\ \text{for } i = 0, 1, 2, \cdots, \\ \quad By_{i+1} = Ax_i, k_{i+1} = \|y_{i+1}\| \\ \quad x_{i+1} = y_{i+1}/k_{i+1}. \end{array} \tag{6.34}$$

We must solve one linear system in each step. In general, Cholesky decomposition of B is necessary.

Inverse power method for $Ax = \lambda Bx$:

$$\begin{array}{l} \text{Given } x_0 \neq 0, \\ \text{for } i = 0, 1, 2, \cdots, \\ \quad \sigma_i = x_i^T A x_i / x_i^T B x_i \\ \quad (A - \sigma_i B) x_{i+1} = k_{i+1} B x_i, \\ \quad k_{i+1} \text{ is chosen so that } \|x_{i+1}\|_B = 1. \end{array} \tag{6.35}$$

Reduction: Let $B = LL^T$. Substitute A by $L^{-1}AL^{-T}$ in (6.33) then we have (here $x_i \leftrightarrow z_i$):

$$\sigma_i = \frac{z_i^T L^{-1} A L^{-T} z_i}{z_i^T z_i} = \frac{x_i^T A x_i}{x_i^T B x_i}, \text{ where } L^{-T} z_i = x_i,$$

and

$$(L^{-1}AL^{-T} - \sigma_i I) z_{i+1} = k_{i+1} z_i \Leftrightarrow (A - \sigma_i B) x_{i+1} = k_{i+1} B x_i.$$

Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of $A - \lambda B$. Then the power iteration (6.34) converges to λ_1. Let $\{\hat{x}_i\}_{i=1}^n$ be the complete system of eigenvectors,i.e.,

$$\hat{x}_i^T B \hat{x}_j = \delta_{ij} \text{ and } \hat{x}_i^T A \hat{x}_j = \lambda_i \delta_{ij} \text{ for all } i, j = 1, \cdots, n.$$

Let $y_1 = \sum_{j=1}^n c_i^1 \hat{x}_i$, $y_k = \sum_{i=1}^n c_i^k \hat{x}_i$. Then it holds

$$y_{k+1} = \sum_{i=1}^n c_i^{k+1} \hat{x}_i = B^{-1} A \sum_{i=1}^n c_i^k \hat{x}_i = \sum_{i=1}^n c_i^k \lambda_i \hat{x}_i.$$

This implies that $c_i^{k+1} = \lambda_i c_i^k$, and thus $c_i^k = \lambda_i^k c_i^1$. Therefore, we have

$$y_k = \lambda_1^k \{c_1^1 \hat{x}_1 + \sum_{\nu=2}^n (\frac{\lambda_\nu}{\lambda_1})^k c_\nu^1 \hat{x}_\nu\}.$$

Normalizing y_k we get that x_k converges to $\hat{x}_1$.

Cost of computation:

Matrix $\times$ vector Ax_i,

Solve the linear system $By_{i+1} = Ax_i$.

Determination of the eigenvalue: $k_{i+1} \to |\lambda_1|$.

Although we have $k_{i+1} \to |\lambda_1|$, the better approximation of λ_1 is $R[x_i]$. Let $x_i = \hat{x}_1 + \epsilon d$, where $d \in span\{\hat{x}_2, \cdots \hat{x}_n\}$ and $\|d\| = 1$. Then

$$\begin{aligned} R[x_i] &= R[\hat{x}_1 + \epsilon d] = \frac{(\hat{x}_1 + \epsilon d)^T A (\hat{x}_1 + \epsilon d)}{(\hat{x}_1 + \epsilon d)^T B (\hat{x}_1 + \epsilon d)} \\ &= \frac{\hat{x}_1^T A \hat{x}_1 + \epsilon^2 d^T A d}{\hat{x}_1^T B \hat{x}_1 + \epsilon^2 d^T B d} = \frac{\hat{x}_1^T A \hat{x}_1}{\hat{x}_1^T B \hat{x}_1} + O(\epsilon^2) \\ &= \lambda_1 + O(\epsilon^2). \end{aligned}$$

Error of eigenvalue $\approx$ (error of eigenvector)^2.

Compute the other eigenvalues and eigenvectors:

Suppose $\lambda_1, \hat{x}_1$ are computed. Power method does not converge to $\lambda_1, \hat{x}_1$, if it satisfies

$$x_i^T B\hat{x}_1 = 0, \quad i = 0, 1, 2, \cdots . \tag{6.36}$$

If we start with x_0 satisfying $x_0^T B\hat{x}_1 = 0$, then all iterate x_i satisfy (6.36) theoretically (since $c_1^1 = 0$). Because of roundoff error we shall perform the reorthogonalization:

$$\begin{aligned} &B\tilde{y}_{i+1} = Ax_i, \\ &y_{i+1} = \tilde{y}_{i+1} - (\hat{x}_i^T B\tilde{y}_{i+1})\hat{x}_1, \\ &x_{i+1} = y_{i+1}/\|y_{i+1}\|_B. \end{aligned}$$

In general: Suppose $\lambda_1, \cdots \lambda_p, \hat{x}_1, \cdots \hat{x}_p$ are computed, then we perofrm the following reorthogonalization:

$$\begin{aligned} &B\tilde{y}_{i+1} = Ax_i, \\ &y_{i+1} = \tilde{y}_{i+1} - \textstyle\sum_{j=1}^{p}(\hat{x}_j^T B\tilde{y}_{i+1})\hat{x}_j, \\ &x_{i+1} = y_{i+1}/\|y_{i+1}\|_B. \end{aligned}$$

Here $x_i^T B\hat{x}_j = 0$, for $j = 1, \cdots, p$, and $i = 0, 1, 2, \cdots$.

Simultaneous vector-iteration:

Determine the p ($p > 1$) largest eigenvalues and the associated eigenvectors of (6.1). Compute simultaneuously the approximations $x_1^{(i)}, \cdots, x_p^{(i)}$. Let

$$X^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)}) \in \mathbf{R}_{n\times p}. \tag{6.37}$$

We demand $X^{(i)}$ satisfies the following relation:

$$X^{(i)^T} BX^{(i)} = I_p. \tag{6.38}$$

Since $\hat{x}_i^T Bx_j = \delta_{ij}$, the columns of $X^{(i)}$ are nearly B-orthogonal. From (6.34) we construct $X^{(i)}$ by

$$BY^{(i)} = AX^{(i-1)} \tag{6.39}$$

and then

$$X^{(i)} = Y^{(i)}C_i, \tag{6.40}$$

where $C_i \in \mathbf{R}^{p\times p}$ and is chosen such that (6.38) is satisfied. We have the following methods for determining C_i .

(a) Apply orthogonalization algorithm to the columns of $Y^{(i)}$, then C_i is an upper triangular matrix.

Let $Y^{(i)} = (y_1^{(i)}, \cdots, y_p^{(i)})$ and $X^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})$.

$$\begin{aligned} &\text{For } k = 1, \cdots, p, \\ &h_k = y_k^{(i)} - \textstyle\sum_{\nu=1}^{k-1}(y_k^{(i)^T} Bx_\nu^{(i)})x_\nu^{(i)}, \\ &x_k^{(i)} = h_k/(h_k^T Bh_k)^{1/2}. \end{aligned}$$

The first column $x_1^{(i)}$ is the same as that we apply power method to $x_1^{(0)}$. Convergence can be slow.

(b) Define $G_i = {Y^{(i)}}^T B Y^{(i)}$, then G_i is positive definite. There exists an orthogonal matrix V_i and $D_i = diag(d_j)$ with $d_1 \geq d_2 \geq \cdots \geq d_p > 0$ such that

$$G_i = V_i D_i V_i^T.$$

Let $X^{(i)} = Y^{(i)} C_i$ where

$$C_i = V_i D_i^{-1/2} \quad \text{and} \quad D_i^{-1/2} = diag(1/\sqrt{d_1}, \cdots, 1/\sqrt{d_p}). \tag{6.41}$$

Check

$$\begin{aligned} {X^{(i)}}^T B X^{(i)} &= C_i^T {Y^{(i)}}^T B Y^{(i)} C_i = (V_i D_i^{-1/2})^T G_i (V_i D_i^{-1/2}) \\ &= D_i^{-1/2} V_i^T G_i V_i D_i^{-1/2} = I_p. \end{aligned}$$

So the columns of $X^{(i)}$ are B-orthogonal. Method (b) brings the approximations in the correct order.

Example: Let $X^{(1)} = (x_2, x_3, x_1)$, where $x_i^T B x_j = \delta_{ij}$ and $A x_i = \lambda_i B x_i, i, j = 1, 2, 3$.
Method (a): $X^{(i)} = X^{(1)}$, $Y^{(2)} = B^{-1} A X^{(1)} = (\lambda_2 x_2, \lambda_3 x_3, \lambda_1 x_1)$. Then

$$X^{(2)} = (x_2, x_3, x_1) = X^{(1)}.$$

Method (b):

$$G_2 = \begin{bmatrix} \lambda_2^2 & 0 & 0 \\ 0 & \lambda_3^2 & 0 \\ 0 & 0 & \lambda_1^2 \end{bmatrix}, \quad D_2 = \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ 0 & \lambda_2^2 & 0 \\ 0 & 0 & \lambda_3^2 \end{bmatrix},$$

$$V_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad C_2 = V_2 D_2^{-1/2} = \begin{bmatrix} 0 & \lambda_2^{-1} & 0 \\ 0 & 0 & \lambda_3^{-1} \\ \lambda_1^{-1} & 0 & 0 \end{bmatrix}.$$

Then

$$X^{(2)} = Y^{(2)} C_2 = (x_1, x_2, x_3).$$

Method (b) forces the eigenvectors in the correct order.

(6.39) and (6.40) imply Treppen iteration (F.L. Bauer 1957) :

For $B = I$:

$$A X^{(i-1)} = Y^{(i)} = X^{(i)} C_i^{-1} = X^{(i)} R_i, \tag{6.42}$$

where R_i is upper triangular.
$p = n$: See the connection with QR Algorithm.

$$A_i = {X^{(i-1)}}^T A X^{(i-1)}, \quad Q_i = {X^{(i-1)}}^T X^{(i)},$$

$$A_i = Q_i R_i, \quad A_{i+1} = R_i Q_i.$$

$p < n$: and $B = LL^T$ positive definite: Treppen iteration for $L^{-1}AL^{-T}$ leads to $Z^{(i)}$:

$$L^{-1}AL^{-T}Z^{(i-1)} = Z^{(i)}R_i\ , \quad {Z^{(i)}}^T Z^{(i)} = I_p. \tag{6.43}$$

Let $X^{(i)} = L^{-T}Z^{(i)}$, rewrite (6.43) to $X^{(i)}$:

$$AX^{(i-1)} = BX^{(i)}R_i \tag{6.44}$$

and

$${X^{(i)}}^T L^T L X^{(i)} = I_p = {X^{(i)}}^T B X^{(i)}.$$

Improvement: $B = I$. Recall

Theorem 6.6.13 *A is real and symmetric, $Q \in \mathbf{R}^{n\times p}$ orthogonal and $S \in \mathbf{R}^{p\times p}$ symmetric, then for an eigenvalue $\lambda_i(S)$ of S there exists an eigenvalue $\lambda_{k_i}(A)$ of A such that*

$$|\lambda_i(S) - \lambda_{k_i}(A)| \le \|AQ - QS\|_2,\ i = 1, \cdots, p.$$

Theorem 6.6.14 *Let $S_0 = Q^T AQ$, then*

$$\|AQ - QS_0\|_{2,F} \le \|AQ - QS\|_{2,F}$$

for all symmetric matrix $S \in R^{p\times p}$.

For given orthogonal matrix $X^{(i)}$, if we construct $S_i = {X^{(i)}}^T AX^{(i)}$, then the eigenvalues of S_i are the optimal approximations to the eigenvalues of A (optimal error estimation). Also good error estimation for eigenvectors. From $S_i z = \mu z$ follows that

$$AX^{(i)}z - \mu X^{(i)}z = (AX^{(i)} - X^{(i)}S_i)z,$$

$$\|A(X^{(i)}z) - \mu(X^{(i)}z)\|_2 \le \|AX^{(i)} - X^{(i)}S_i\|_2\|z\|_2.$$

So $X^{(i)}z$ is a good approximation to an eigenvector of A.

B =positive definite:

Given $n \times p$ matrix S with $rank(S) = p$. Let $\mathcal{S} = span(S)$. Find a new base of $\mathcal{S}$, which presents a good approximation to eigenvectors of

$$Ax = \lambda Bx.$$

(6.1) is equivalent to :

$$\hat{A}\hat{x} = \lambda\hat{x} \text{ with } \hat{A} = B^{-1/2}AB^{-1/2},\ \hat{x} = B^{1/2}x. \tag{6.45}$$

Orthonormalize $B^{1/2}S(S \to B^{1/2}S)$ and results

$$\hat{S} = B^{1/2}S(S^T BS)^{-1/2}. \tag{6.46}$$

(Check$\hat{S}^T\hat{S} = I_p$). From above we know that the eigenvalue μ_i of $\hat{H} = \hat{S}^T\hat{A}\hat{S}$ are a good approximation to an eigenvalue of (6.45), so of (6.1) and $\hat{g}_i$ is the associated

eigenvector, then $\hat{S}\hat{g}_i$ is a good approximation to an eigenvector of (6.45), so $B^{-1/2}\hat{S}\hat{g}_i$ an approximation to an eigenvector of (6.1). Rewrite A, B, S:

$$\begin{aligned}\hat{H} &= (S^TBS)^{-1/2}S^TB^{1/2}B^{-1/2}AB^{-1/2}B^{1/2}S(S^TBS)^{-1/2} \\ &= (S^TBS)^{-1/2}(S^TAS)(S^TBS)^{-1/2}\end{aligned}$$

Then $\hat{H}\hat{g}_i$ corresponds to

$$(S^TAS - \mu_i S^TBS)\underbrace{(S^TBS)^{-1/2}\hat{g}_i}_{g_i} = 0,$$

i.e.

$$(A_s - \mu_i B_s)g_i = 0 \quad \text{with} \quad \begin{cases} A_s = S^TAS, \\ B_s = S^TBS. \end{cases} \tag{6.47}$$

If S is given, construct A_s, B_s. The eigenvalues of $A_s z = \mu B_s z$ are good approximations to eigenvalues of (6.1). Compute the eigenvectors g_i of (6.47), then Sg_i are approximations to the eigenvectors of (6.1).

Some variant simultaneuous vector iterations ($B = I$):

(a)

$$\begin{array}{l}(1)\ Y^{(\nu)} = AX^{(\nu-1)}, \\ (2)\ \text{Orthonormalize}\ \ Y^{(\nu)} = Q_\nu R_\nu\ (QR\ \text{decomposition}), \\ (3)\ \text{Compute}\ \ H_\nu = Q_\nu^T A Q_\nu, \\ (4)\ \text{Solve the complete eigenvalue system for}\ \ H_\nu, \\ \quad H_\nu = G_\nu \Theta_\nu G_\nu^T,\ \ G_\nu : \text{orthogonal and } \Theta_\nu : \text{diagonal}, \\ (5)\ X^{(\nu)} = Q_\nu G_\nu\ (\text{The element of } \Theta_\nu \text{ are in decreasing order}).\end{array} \tag{6.48}$$

The computation of (1) and (3) are expansive, it can be avoided by the following way. Since the invariant subspaces and eigenvectors of A and A^{-2} are equal, so we can consider the matrix A^{-2} instead of A. The eigenvectors of $Q_\nu^T A^{-2} Q_\nu$ are the good approximations for the eigenvectors of A.

Compute

$$\begin{aligned}Q_\nu^T A^{-2} Q_\nu &= (R_\nu^{-1})^T \overbrace{X^{(\nu-1)^T} A A^{-2} A X^{(\nu-1)}}^{I_p} R_\nu^{-1}. \\ &\quad (\text{Here } Q_\nu = AX^{(\nu-1)}R_\nu^{-1} \text{ from (1) (2) above}) \\ &= R_\nu^{-T} R_\nu^{-1} = (R_\nu R_\nu^T)^{-1}.\end{aligned}$$

So we have the following new method:
(b)

$$\begin{array}{l}(1)\ Y^{(\nu)} = AX^{(\nu-1)}, \\ (2)\ \text{Orthonormalize}\ \ Y^{(\nu)} = Q_\nu R_\nu, \\ (3)\ \text{Compute}\ \ \tilde{H}_\nu = R_\nu R_\nu^T, \\ (4)\ \text{Solve}\ \ \tilde{H}_\nu = P_\nu \Delta_\nu^2 P_\nu^T\ ,\ \ P_\nu : \text{orthogonal},\ \Delta_\nu : \text{diagonal}, \\ (5)\ X^{(\nu)} = Q_\nu P_\nu.\end{array} \tag{6.49}$$

(c) Third variant compution of Q_ν:

Find F_ν such that $Y^{(\nu)}F_\nu$ is orthogonal. So

$$F_\nu^T {Y^{(\nu)}}^T Y^{(\nu)} F_\nu = I \tag{6.50}$$

and

$${Y^{(\nu)}}^T Y^{(\nu)} = F_\nu^{-T} F_\nu^{-1} = (F_\nu F_\nu^T)^{-1}.$$

On the other hand $Y^{(\nu)}F_\nu$ diagonalize A^{-2}, i.e.,

$$F_\nu^T Y^{(\nu)} A^{-2} {Y^{(\nu)}}^T F_\nu = \Delta_\nu^{-2} \text{ diagonal.} \tag{6.51}$$

From (6.51) and because of $Y^{(\nu)} = AX^{(\nu-1)}$ follows

$$\Delta_\nu^{-2} = F_\nu^T \underbrace{{X^{(\nu-1)}}^T A A^{-2} A X^{(\nu-1)}}_{I_p} F_\nu = F_\nu^T F_\nu.$$

Thus $I = \Delta_\nu F_\nu^T F_\nu \Delta_\nu$ and then $F_\nu \Delta_\nu$ is orthogonal. Using (6.50), we have

$$\begin{aligned} H_\nu = {Y^{(\nu)}}^T Y^{(\nu)} &= (F_\nu^{-T}\Delta_\nu^{-1})\Delta_\nu^2(\Delta_\nu^{-1}F_\nu^{-1}) \\ &= (\underbrace{F_\nu\Delta_\nu}_{\text{ortho.}})^{-T} \underbrace{\Delta_\nu^2}_{\text{diag.}} (\underbrace{F_\nu\Delta_\nu}_{\text{ortho.}})^{-1}. \end{aligned}$$

The diagonal elements of Δ_ν^2 are the eigenvalues of H_ν and the column of $F_\nu\Delta_\nu$ are the eigenvectors of H_ν, therefore we can compute F_ν as follows:

$$\begin{aligned} &(1) Y^{(\nu)} = AX^{(\nu-1)}, \\ &(2) \text{Compute } \hat{H}_\nu = {Y^{(\nu)}}^T Y^{(\nu)}, \\ &(3) \text{Compute } \hat{H}_\nu = B_\nu \Delta_\nu^2 B_\nu^T \text{ complete eigensystem of } \hat{H}_\nu, \\ &(4) X^{(\nu)} = Y^{(\nu)} B_\nu \Delta_\nu^{-1} (= Y^{(\nu)} F_\nu). \end{aligned} \tag{6.52}$$

The cost of computation of (6.52) is more favorable than of (6.49).

Chapter 7

Lanczos Methods

In this chapter we develop the Lanczos method, a technique that is applicable to large sparse, symmetric eigenproblems. The method involves tridiagonalizing the given matrix A. However, unlike the Householder approach, no intermediate (an full) submatrices are generated. Equally important, information about $A's$ extremal eigenvalues tends to emerge long before the tridiagonalization is complete. This makes the Lanczos algorithm particularly useful in situations where a few of $A's$ largest or smallest eigenvalues are desired.

7.1 The Lanczos Algorithm

Suppose $A \in \mathbf{R}^{n\times n}$ is large, sparse and symmetric. There exists an orthogonal matrix Q, which transforms A to a tridiagonal matrix T.

$$Q^T AQ = T \equiv \text{ tridiagonal.} \tag{1.1}$$

Remark
(1) Such Q can be generated by Householder transformations or Givens rotations.
(2) Almost for all A (i.e. all eigenvalues are distinct) and almost for any $q_1 \in \mathbf{R^n}$ with $\|q_1\|_2 = 1$, there exists an orthogonal matrix Q with first column q_1 satisfying (1.1). q_1 determines T uniquely up to the sign of the columns (that is, we can multiply each column with -1).

Let ($x \in \mathbf{R}^n$)

$$K[x, A, m] = [x, Ax, A^2x, \cdots, A^{m-1}x] \in \mathbf{R}^{n\times m}. \tag{1.2}$$

$K[x, A, m]$ is called a Krylov-matrix.

Let

$$\mathcal{K}(x, A, m) = Range(K[x, A, m]) = Span(x, Ax, \cdots, A^{m-1}x). \tag{1.3}$$

$\mathcal{K}(x, A, m)$ is called the Krylov-subspace generated by $K[x, A, m]$.

Remark: For each $H \in \mathbf{C}^{n\times m}$ or $\mathbf{R}^{n\times m}$ ($m \le n$) with $rank(H) = m$, there exists an $Q \in \mathbf{C}^{n\times m}$ or $\mathbf{R}^{n\times m}$ and an upper triangular $R \in \mathbf{C}^{m\times m}$ or $\mathbf{R}^{m\times m}$ with $Q^*Q = I_m$ such that

$$H = QR. \tag{1.4}$$

Q is uniquely determined, if we require all $r_{ii} > 0$.

Theorem 7.1.1 *Let A be symmetric (Hermitian), $1 \le m \le n$ be given and dim $\mathcal{K}(x, A, m) = m$ then*
(a) If

$$K[x, A, m] = Q_m R_m \tag{1.5}$$

is an QR decomposition, then $Q_m^* A Q_m = T_m$ is an $m \times m$ tridiagonal matrix and satisfies

$$AQ_m = Q_m T_m + r_m e_m^T, \quad Q_m^* r_m = 0. \tag{1.6}$$

(b) Let $\|x\|_2 = 1$. If $Q_m \in C^{n\times m}$ with the first column x and $Q_m^* Q_m = I_m$ and satisfies

$$AQ_m = Q_m T_m + r_m e_m^T,$$

where T_m is tridiagonal, then

$$K[x, A, m] = [x, Ax, \cdots, A^{m-1}x] = Q_m[e_1, T_m e_1, \cdots, T_m^{m-1} e_1] \tag{1.7}$$

is an QR decomposition of $K[x, A, m]$.

Proof: (a) Since

$$A\mathcal{K}(x, A, j) \subset \mathcal{K}(x, A, j+1), \quad j < m. \tag{1.8}$$

From (1.5) we have

$$Span(q_1, \cdots, q_i) = \mathcal{K}(x, A, i), \quad i \le m. \tag{1.9}$$

So we have

$$q_{i+1} \perp \mathcal{K}(x, A, i) \overset{(1.8)}{\supset} A\mathcal{K}(x, A, i-1) = A(span(q_1, \cdots, q_{i-1})).$$

This implies

$$q_{i+1}^* A q_j = 0, \quad j = 1, \cdots, i-1, \quad i+1 \le m.$$

That is

$$(Q_m^* A Q_m)_{ij} = (T_m)_{ij} = q_i^* A q_j = 0 \text{ for } i > j+1.$$

So T_m is upper Hessenberg and then tridiagonal (since T_m is Hermitian).

It remains to show (1.6). Since

$$[x, Ax, \cdots, A^{m-1}x] = Q_m R_m$$

and

$$AK[x, A, m] = K[x, A, m] \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} + A^m x e_m^T,$$

we have

$$AQ_m R_m = Q_m R_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} + Q_m Q_m^* A^m x e_m^T + (I - Q_m Q_m^*) A^m x e_m^T.$$

Then

$$\begin{aligned} AQ_m &= Q_m[R_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} + Q_m^* A^m x e_m^T] R_m^{-1} + (I - Q_m Q_m^*) A^m x e_m^T R_m^{-1} \\ &= Q_m[R_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} R_m^{-1} + \gamma Q_m^* A^m x e_m^T] + \underbrace{\gamma(I - Q_m Q_m^*) A^m x}_{r_m} e_m^T \\ &= Q_m H_m + r_m e_m^T \text{ with } Q_m^* r_m = 0, \end{aligned}$$

where H_m is an upper Hessenberg matrix. But $Q_m^* A Q_m = H_m$ is Hermitian, so $H_m = T_m$ is tridiagonal.

(b) We check (1.7):

$x = Q_m e_1$ coincides the first column. Suppose that i-th columns are equal, i.e.

$$\begin{aligned} A^{i-1}x &= Q_m T_m^{i-1} e_1 \\ A^i x &= A Q_m T_m^{i-1} e_1 \\ &= (Q_m T_m + r_m e_m^T) T_m^{i-1} e_1 \\ &= Q_m T_m^i e_1 + r_m e_m^T T_m^{i-1} e_1. \end{aligned}$$

But $e_m^T T_m^{i-1} e_1 = 0$ for $i < m$. Therefore, $A^i x = Q_m T^i e_1$ the $(i+1)$-th columns are equal. It is clearly that $(e_1, T_m e_1, \cdots, T_m^{m-1} e_1)$ is an upper triangular matrix. ■

Theorem 7.1.2 *If $x = q_1$ with $\|q_1\|_2 = 1$ satisfies*

$$rank(K[x, A, n]) = n$$

(that is $\{x, Ax, \cdots, A^{n-1}x\}$ are linearly independent), then there exists an unitary matrix Q with first column q_1 such that $Q^*AQ = T$ is tridiagonal.

Proof: From Theorem 7.1.1(a) $m = n$ we have $Q_m = Q$ unitary and $AQ = QT$.

Uniqueness: Let $Q^*AQ = T$, $\quad \tilde{Q}^* A \tilde{Q} = \tilde{T}$ and $Q_1 e_1 = \tilde{Q} e_1$

$$\begin{aligned} &\Rightarrow \quad K[q_1, A, n] = QR = \tilde{Q}\tilde{R} \\ &\Rightarrow \quad Q = \tilde{Q}D, \quad R = D\tilde{R}. \end{aligned}$$

Substitute Q by QD, where $D = diag(\epsilon_1, \cdots, \epsilon_n)$ with $|\epsilon_i| = 1$. Then

$$(QD)^* A (QD) = D^* Q^* A Q D = D^* T D = \text{ tridiagonal.}$$

So Q is unique up to multiplying the columns of Q by a factor ϵ with $|\epsilon| = 1$. ■

In the following paragraph we will investigate the Lanczos algorithm for the real case, i.e., $A \in \mathbf{R}^{n \times n}$.

How to find an orthogonal matrix $Q = (q_1, \cdots, q_n)$ with $Q^TQ = I_n$ such that $Q^TAQ = T$ = tridiagonal and Q is almost uniquely determined. Let

$$AQ = QT, \tag{1.10}$$

$$Q = [q_1, \cdots, q_n] \text{ and } T = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

It implies that the j-th column of (1.10) forms:

$$Aq_j = \beta_{j-1}q_{j-1} + \alpha_j q_j + \beta_j q_{j+1}, \tag{1.11}$$

for $j = 1, \cdots, n$ with $\beta_0 = \beta_n = 0$. By multiplying (1.11) by q_j^T we obtain

$$q_j^T Aq_j = \alpha_j. \tag{1.12}$$

Define $r_j = (A - \alpha_j I)q_j - \beta_{j-1}q_{j-1}$. Then

$$r_j = \beta_j q_{j+1}$$

with

$$\beta_j = \pm\|r_j\|_2 \tag{1.13}$$

and if $\beta_j \neq 0$ then

$$q_{j+1} = r_j/\beta_j. \tag{1.14}$$

So we can determine the unknown α_j, β_j, q_j in the following order:

$$\text{Given } q_1, \alpha_1, r_1, \beta_1, q_2, \alpha_2, r_2\beta_2, q_3, \cdots.$$

The above formula define the Lanczos iterations:

$$\begin{array}{l} j = 0,\ r_0 = q_1\ ,\ \beta_0 = 1\ ,\ q_0 = 0 \\ \text{Do while } (\beta_j \neq 0) \\ \quad q_{j+1} = r_j/\beta_j\ ,\ j := j+1 \\ \quad \alpha_j = q_j^T Aq_j\ , \\ \quad r_j = (A - \alpha_j I)q_j - \beta_{j-1}q_{j-1}, \\ \quad \beta_j = \|r_j\|_2. \end{array} \tag{1.15}$$

There is no loss of generality in choosing the β_j to be positive. The q_j are called Lanczos vectors. With careful overwriting and use of the formula $\alpha_j = q_j^T(Aq_j - \beta_{j-1}q_{j-1})$, the whole process can be implemented with only a pair of n-vectors.

Algorithm 7.1.1 (Lanczos Algorithm):
Given a symmetric $A \in \mathbf{R}^{n\times n}$ and $w \in \mathbf{R}^n$ having unit 2-norm. The following algorithm computes a $j \times j$ symmetric tridiagonal matrix T_j with the property that $\sigma(T_j) \subset \sigma(A)$.

The diagonal and subdiagonal elements of T_j are stored in $\alpha_1, \cdots, \alpha_j$ and $\beta_1, \cdots, \beta_{j-1}$ respectively.

$v_i := 0 \quad (i = 1, \cdots, n)$
$\beta_0 := 1$
$j := 0$
Do while $(\beta_j \neq 0)$
if $(j \neq 0)$,
then for $i = 1, \cdots, n$,
$t := w_i, w_i := v_i/\beta_j, v_i := -\beta_j t.$
$v := Aw + v,$
$j := j + 1,$
$\alpha_j := w^T v,$
$v := v - \alpha_j w,$
$\beta_j := \|v\|_2.$

Remark
(1) If the sparity is exploited and only kn flops are involved in each call (Aw) $(k \ll n)$, then each Lanczos step requires about $(4+k)n$ flops to execute.
(2) The iteration stops before complete tridiagonalizaton if q_1 is contained in a proper invariant subspace. From the iteration (1.15) we have

$$A(q_1, \cdots, q_m) = (q_1, \cdots, q_m) \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \beta_{m-1} \\ & \ddots & \ddots & \\ & & \beta_{m-1} & \alpha_m \end{bmatrix} + \underbrace{(0, \cdots, 0, \overbrace{\beta_m q_{m+1}}^{r_m})}_{r_m e_m^T}$$

$$\beta_m = 0 \text{ if and only if } r_m = 0.$$

This implies

$$A(q_1, \cdots, q_m) = (q_1, \cdots, q_m) T_m.$$

That is

$$Range(q_1, \cdots, q_m) = Range(K[q_1, A, m])$$

is the invariant subspace of A and the eigenvalues of T_m are the eigenvalues of A.

Theorem 7.1.3 *Let A be symmetric and q_1 be a given vector with $\|q_1\|_2 = 1$. The Lanczos iterations (1.15) runs until $j = m$ where $m = rank[q_1, Aq_1, \cdots, A^{n-1}q_1]$. Moreover, for $j = 1, \cdots, m$ we have*

$$AQ_j = Q_j T_j + r_j e_j^T \tag{1.16}$$

with

$$T_j = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ & & \beta_{j-1} & \alpha_j \end{bmatrix} \text{ and } Q_j = [q_1, \cdots, q_j]$$

has orthonormal columns satisfying $Range(Q_j) = \mathcal{K}(q_1, A, j)$.

Proof: By induction on j. Suppose the iteration has produced $Q_j = [q_1, \cdots, q_j]$ such that $Range(Q_j) = \mathcal{K}(q_1, A, j)$ and $Q_j^T Q_j = I_j$. It is easy to see from (1.15) that (1.16) holds. Thus

$$Q_j^T A Q_j = T_j + Q_j^T r_j e_j^T.$$

Since $\alpha_i = q_i^T A q_i$ for $i = 1, \cdots, j$ and

$$q_{i+1}^T A q_i = q_{i+1}^T(\beta_i q_{i+1} + \alpha_i q_i + \beta_{i-1} q_{i-1}) = q_{i+1}^T(\beta_i q_{i+1}) = \beta_i$$

for $i = 1, \cdots, j-1$ we have $Q_j^T A Q_j = T_j$. Consequently $Q_j^T r_j = 0$.

If $r_j \neq 0$ then $q_{j+1} = r_j / \|r_j\|_2$ is orthogonal to $q_1, \cdots, q_j$ and

$$q_{j+1} \in Span\{Aq_j, q_j, q_{j-1}\} \subset \mathcal{K}(q_1, A, j+1).$$

Thus $Q_{j+1}^T Q_{j+1} = I_{j+1}$ and $Range(Q_{j+1}) = \mathcal{K}(q_1, A, j+1)$.

On the other hand, if $r_j = 0$, then $AQ_j = Q_j T_j$. This says that $Range(Q_j) = \mathcal{K}(q_1, A, j)$ is invariant. From this we conclude that $j = m = dim[\mathcal{K}(q_1, A, n)]$. ■

Encountering a zero β_j in the Lanczos iteration is a welcome event in that it signals the computation of an exact invariant subspace. However an exactly zero or even small β_j is rarly in practice. Consequently, other explanations for the convergence of $T_j's$ eigenvalues must be sought.

Theorem 7.1.4 *Suppose that j steps of the Lanczos algorithm have been performed and that*

$$S_j^T T_j S_j = diag(\theta_1, \cdots, \theta_j)$$

is the Schur decomposition of the tridiagonal matrix T_j, if $Y_j \in \mathbf{R}^{n \times j}$ is defined by

$$Y_j = [y_1, \cdots, y_j] = Q_j S_j$$

then for $i = 1, \cdots, j$ we have

$$\|Ay_i - \theta_i y_i\|_2 = |\beta_j||s_{ji}|$$

where $S_j = (s_{pq})$.

Proof: Post-multiplying (1.16) by S_j gives

$$AY_j = Y_j diag(\theta_1, \cdots, \theta_j) + r_j e_j^T S_j,$$

i.e.,

$$Ay_i = \theta_i y_i + r_j(e_j^T S_j e_i) \ , \quad i = 1, \cdots, j.$$

The proof is complete by taking norms and recalling $\|r_j\|_2 = |\beta_j|$ ■.

Remark: The theorem provides error bounds for $T_j's$ eigenvalues:

$$\min_{\mu \in \sigma(A)} |\theta_i - \mu| \leq |\beta_j||s_{ji}| \quad i = 1, \cdots, j.$$

Note that in section 10 the (θ_i, y_i) are Ritz pairs for the subspace $R(Q_j)$.

If we use the Lanczos method to compute $AQ_j = Q_jT_j + r_je_j^T$ and set $E = \tau ww^T$ where $\tau = \pm 1$ and $w = aq_j + br_j$, then it can be shown that

$$(A + E)Q_j = Q_j(T_j + \tau a^2 e_j e_j^T) + (1 + \tau ab)r_j e_j^T.$$

If $0 = 1 + \tau ab$, then the eigenvalues of the tridiagonal matrix

$$\tilde{T}_j = T_j + \tau a^2 e_j e_j^T$$

are also eigenvalues of $A+E$. We may then conclude from theorem 6.1.2 that the interval $[\lambda_i(T_j), \lambda_{i-1}(T_j)]$ where $i = 2, \cdots, j$, each contains an eigenvalue of $A + E$.

Suppose we have an approximate eigenvalue $\tilde{\lambda}$ of A. One possibility is to choose τa^2 so that

$$det(\tilde{T}_j - \tilde{\lambda} I_j) = (\alpha_j + \tau a^2 - \tilde{\lambda})p_{j-1}(\tilde{\lambda}) - \beta_{j-1}^2 p_{j-2}(\tilde{\lambda}) = 0,$$

where the polynomial $p_i(x) = det(T_i - xI_i)$ can be evaluated at $\tilde{\lambda}$ using (5.3).

The following theorems are known as the Kaniel-Paige theory for the estimation of eigenvalues which obtained via the Lanczos algorithm.

Theorem 7.1.5 *Let A be $n \times n$ symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding orthonormal eigenvectors $z_1, \cdots, z_n$. If $\theta_1 \geq \cdots \geq \theta_j$ are the eigenvalues of T_j obtained after j steps of the Lanczos iteration, then*

$$\lambda_1 \geq \theta_1 \geq \lambda_1 - \frac{(\lambda_1 - \lambda_n)\tan(\phi_1)^2}{[c_{j-1}(1 + 2\rho_1)]^2},$$

where $\cos\phi_1 = |q_1^T z_1|$, $\rho_1 = (\lambda_1 - \lambda_2)/(\lambda_2 - \lambda_n)$ and c_{j-1} is the Chebychev polynomial of degree $j - 1$.

Proof: From Courant-Fischer theorem we have

$$\theta_1 = \max_{y \neq 0} \frac{y^T T_j y}{y^T y} = \max_{y \neq 0} \frac{(Q_j y)^T A (Q_j y)}{(Q_j y)^T (Q_j y)} = \max_{0 \neq w \in \mathcal{K}(q_1, A, j)} \frac{w^T A w}{w^T w}.$$

Since λ_1 is the maximum of $w^T Aw/w^T w$ over all nonzero w, it follows that $\lambda_1 \geq \theta_1$. To obtain the lower bound for θ_1, note that

$$\theta_1 = \max_{p \in P_{j-1}} \frac{q_1^T p(A) A p(A) q_1}{q_1^T p(A)^2 q_1},$$

where P_{j-1} is the set of all $j - 1$ degree polynomials. If

$$q_1 = \sum_{i=1}^{n} d_i z_i$$

then

$$\frac{q_1^T p(A) A p(A) q_1}{q_1^T p(A)^2 q_1} = \frac{\sum_{i=1}^n d_i^2 p(\lambda_i)^2 \lambda_i}{\sum_{i=1}^n d_i^2 p(\lambda_i)^2}$$

$$\geq \lambda_1 - (\lambda_1 - \lambda_n)\frac{\sum_{i=2}^{n} d_i^2 p(\lambda_i)^2}{d_1^2 p(\lambda_1)^2 + \sum_{i=2}^{n} d_i^2 p(\lambda_i)^2}.$$

We can make the lower bound tight by selecting a polynomial $p(x)$ that is large at $x = \lambda_1$ in comparison to its value at the remaining eigenvalues. Set

$$p(x) = c_{j-1}[-1 + 2\frac{x - \lambda_n}{\lambda_2 - \lambda_n}],$$

where $c_{j-1}(z)$ is the $(j-1)$-th Chebychev polynomial generated by

$$c_j(z) = 2zc_{j-1}(z) - c_{j-2}(z), \quad c_0 = 1, c_1 = z.$$

These polynomials are bounded by unity on [-1,1]. It follows that $|p(\lambda_i)|$ is bounded by unity for $i = 2, \cdots, n$ while $p(\lambda_1) = c_{j-1}(1 + 2\rho_1)$. Thus,

$$\theta_1 \geq \lambda_1 - (\lambda_1 - \lambda_n)\frac{(1 - d_1^2)}{d_1^2}\frac{1}{c_{j-1}^2(1 + 2\rho_1)}.$$

The desired lower bound is obtained by noting that $\tan(\phi_1)^2 = (1 - d_1^2)/d_1^2$. ■

Corollary 7.1.6 *Using the same notation as the theorem 7.1.5*

$$\lambda_n \leq \theta_j \leq \lambda_n + \frac{(\lambda_1 - \lambda_n)\tan^2(\phi_n)}{c_{j-1}^2(1 + 2\rho_n)},$$

where $\rho_n = (\lambda_{n-1} - \lambda_n)/(\lambda_1 - \lambda_{n-1})$ *and* $\cos(\phi_n) = |q_1^T z_n|$.

Proof: Apply theorem 7.1.5 with A replaced by $-A$. ■

Example:

$$\begin{aligned} L_{j-1} &\equiv \frac{1}{[C_{j-1}(2\frac{\lambda_1}{\lambda_2} - 1)]^2} \geq \frac{1}{[C_{j-1}(1 + 2\rho_1)]^2} \\ R_{j-1} &= (\frac{\lambda_2}{\lambda_1})^{2(j-1)} \quad \text{power methed} \end{aligned}$$

λ_1/λ_2	j=5	j=25	
1.5	$1.1 \times 10^{-4}/3.9 \times 10^{-2}$	$1.4 \times 10^{-27}/3.5 \times 10^{-9}$	L_{j-1}/R_{j-1}
1.01	$5.6 \times 10^{-1}/9.2 \times 10^{-1}$	$2.8 \times 10^{-4}/6.2 \times 10^{-1}$	L_{j-1}/R_{j-1}

Rounding errors greatly affect the behavior of algorithm 7.1.1, the Lanczos iteration. The basic difficulty is caused by loss of orthogonality among the Lanczos vectors. To avoid these difficulties we can reorthogonalize the Lanczos vectors.

(1) Complete reorthogonalization:
Orthogonalize q_j to all $q_1, \cdots, q_{j-1}$ by

$$q_j := q_j - \sum_{i=1}^{j-1}(q_j^T q_i)q_i.$$

If we incorporate the Householder computations into the Lanczos process, we can produce Lanczos vectors that are orthogonal to working accuracy:

$r_0 := q_1$ (given unit vector)
Determine $P_0 = I - 2v_0v_0^T/v_0^Tv_0$ so $P_0r_0 = e_1$
  $\alpha_1 := q_1^TAq_1$
  Do $j = 1, \cdots, n-1$,
    $r_j := (A - \alpha_j)q_j - \beta_{j-1}q_{j-1}(\beta_0q_0 \equiv 0)$,
    $w := (P_{j-1} \cdots P_0)r_j$.
    Determine $P_j = I - 2v_jv_j^T/v_j^Tv_j$ such that
    $P_jw = (w_1, \cdots, w_j, \beta_j, 0, \cdots, 0)^T$.
    $q_{j+1} := (P_0 \cdots P_j)e_{j+1}$,
    $\alpha_{j+1} := q_{j+1}^TAq_{j+1}$.

This is the complete reorthogonalization Lanczos scheme.

(2) Selective reorthogonalization:

A remarkable, somewhat ironic consequence of the Paige (1971) error analysis is that loss of orthogonality goes hand in hand with convergence of a Ritz pair.

For details of (1) and (2) see the books:
Parlett: "Symmetric Eigenvalue problem" (1980) pp.257–
Golub & Van Loan: "Matrix computation" (1981) pp.332–

Theorem 7.1.7 (Paige Theorem) *Let* $y_i = Q_jS_i$, $i = 1, \ldots, j$ (Ritz vector). *Then*

$$y_i^Tq_{j+1} = r_{ii}/\beta_{ji} \equiv r_{ii}/(\beta_js_{ji}),$$

where $r_{ii} = O(\varepsilon)$, $\sum$(*round-off-error*).

Recall that $\|Ay_i - \theta_iy_i\|_2 = |\beta_j||s_{ji}| \equiv |\beta_{ji}| \leq O(\varepsilon)$ (very small),
$y_j \in \text{span}(Q_j)$,
$q_{j+1}^Ty_i = \frac{r_{ii}}{\beta_{ji}} \approx O(\varepsilon)$ (very small)

$$= \begin{cases} O(\varepsilon) & \text{yes!} \quad \text{if } |\beta_{ji}| = O(1) \\ O(1) & \text{no!} \quad \text{if } |\beta_{ji}| = O(\varepsilon). \end{cases}$$

Loss of orthogonality !

(1) Selective Reorthogonalization:
Select "good" Ritz vectors ($|\beta_{ji}| \approx o(\sqrt{\varepsilon})$) and do reorthogonalization
(ie. $q_{i+1} \perp$ "good" Ritz vector).

(2) Restart: – full reorthogonalization.
– Restart in m-steps ($m = 30 \sim 50$).

Lanczos method

$$
\begin{array}{l}
\text{Given } q_1 \neq 0 \\
\left[\begin{array}{l}
j = 1, 2, ..., m \\
\alpha_j = q_j^T A q_j \\
r_j = (A - \alpha_j I) q_j + \beta_{j-1} q_{j-1} \\
\beta_j = \|r_j\|_2, \text{ if } \beta_j \neq 0, \text{ otherwise stop .} \\
q_{j+1} = r_j / \beta_j \\
\text{end m}
\end{array}\right.
\end{array}
$$

$$A(Q_m) = Q_m T_m + r_m e_m^T,$$

where

$$Q_m^T Q_m = I_m, \quad T_m = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ 0 & & \beta_{m-1} & \alpha_m \end{bmatrix}.$$

$$s_m^T T_m s_m = \Theta_m = \begin{bmatrix} \theta_1 & & 0 \\ & \ddots & \\ 0 & & \theta_m \end{bmatrix}, \text{ Ritz value}$$

$$\|A(Q_m s_j) - Q_j(Q_m s_j)\|_2 = \beta_{jm} \equiv |\beta_j||s_{jm}|,$$

$$S = [s_1, \cdots, s_m], \ j = 1, \ldots, m.$$

Paige Theorem

Since $AQ_j = Q_j T_j + r_j e_j^T$, let

$$AQ_j - Q_j T_j = r_j e_j^T + F_j \tag{1.17}$$

$$I - Q_j^T Q_j = C_j^T + \Delta_j + C_j, \tag{1.18}$$

where C_j is strictly upper triangular and Δ_j is diagonal.
(For simplicity, suppose $(C_j)_{i,i+1} = 0$ and $\Delta_i = 0$.)
Ritz vector $\quad y_i \equiv Q_j s_i$.

Ritz value $\quad s_j^T T_j s_j = \begin{bmatrix} \theta_1 & & 0 \\ & \ddots & \\ 0 & & \theta_j \end{bmatrix} \equiv \Theta_j \quad (T_j s_j = \theta_j s_j).$

Theorem 7.1.8 (Paige Theorem) *Assume* (a) S_j *and* θ_j *are exact !* ($\because j \ll n$)
(b) *local orthogonality is maintained.* (*i.e.* $q_{i+1}^T q_i = 0$, $i = 1, \ldots, j-1$, $r_j^T q_j = 0$, *and* $(C_j)_{i,i+1} = 0$). *Let*

$$F_j^T Q_j - Q_j^T F_j = K_j - K_j^T,$$

$$\Delta_j T_j - T_j \Delta_j \equiv N_j - N_j^T,$$

$$G_j = S_j^T (K_j + N_j) S_j \equiv (r_{ik}).$$

Then

(1) $y_i^T q_{j+1} = r_{ii}/\beta_{ji}$, *where* $y_i = Q_j S_i$, $\beta_{ji} = \beta_j S_{ji}$. $(\star)$

(2) *For* $i \neq k$,

$$(\theta_i - \theta_k) y_i^T y_k = r_{ii}(\frac{S_{jk}}{S_{ji}}) - r_{kk}(\frac{S_{ji}}{S_{jk}}) - (r_{ik} - r_{ki}). \tag{1.19}$$

Proof: From (1.17), $AQ_j - Q_j T_j = r_j e_j^T + F_j$.
(1.17) is multiplied from left by $Q_j^T \Rightarrow$

$$Q_j^T A Q_j - Q_j^T Q_j T_j = Q_j^T r_j e_j^T + Q_j^T F_j. \tag{1.20}$$

$(9.2.7)^T \Rightarrow Q_j^T A^T Q_j - T_j Q_j Q_j^T = e_j r_j^T Q_j + F_j^T Q_j$

$$\begin{aligned}
&(Q_j^T \gamma_j) e_j^T - e_j (Q_j^T \gamma_j)^T \\
=&(C_j^T T_j - T_j C_j^T) + (C_j T_j - T_j C_j) + (\Delta_j T_j - T_j \Delta_j) + F_j^T Q_j - Q_j F_j^T \\
=&(C_j^T T_j - T_j C_j^T) + (C_j T_j - T_j C_j) + (N_j - N_j^T) + (K_j - K_j^T)
\end{aligned}$$

$$\Rightarrow (Q_j^T r_j) e_j^T = C_j T_j - T_j C_j + N_j + K_j \qquad (\star\star)$$

$S_i^T \times (\star\star) \times S_i$ gives

$$\begin{aligned}
y_i^T q_{j+1} \beta_{ji} &= S_i^T (C_j T_j - T_j C_j) S_i + S_i^T (N_j + K_j) S_i \\
&= (S_i^T C_j S_i)\theta_i - \theta_i (S_i^T C_j S_i) + r_{ii} \\
\Rightarrow y_i^T q_{j+1} &= \tfrac{r_{ii}}{\beta_{ji}}
\end{aligned}$$

(9.2.6) can be obtained by $S_i^T \times (9.2.7) \times S_k$.

Remark: To $(\star)$: $y_i^T q_{j+1} = \frac{r_{ii}}{\beta_{ji}}$, $i = 1, \ldots, j$.

$$y_i^T q_{j+1} = \frac{r_{ii}}{\beta_{ji}} = \begin{cases} O(esp), if|\beta_{ji}| = O(1)(\text{not converge!}) \\ O(1), if|\beta_{ji}| = O(esp)(\text{converge for}(\theta_j, y_j)) \end{cases} \tag{1.21}$$

$q_{j+1}^T y_i = O(1)$, q_{j+1} is not orthogonal to $< Q_j >$, $Q_j S_j = y_j$

(1) Full Reorthogonalization by MGS

$$q_{j+1} \perp q_1, \ldots, q_j$$

$$q_{j+1} := q_{j+1} - \sum_{i=1}^{j} (q_{j+1}^T q_i) q_i.$$

(2) Selective Reorthogonalization by MGS
If $|\beta_{ji}| = O(\sqrt{eps})$, (θ_j, y_j) "good" Ritz pair
Do $q_{j+1} \perp q_1, \ldots, q_j$
Else not to do Reorthogonalization

(3) Restart after m-steps
(Do full Reorthogonalization)

(4) Partial Reorthogonalization
Do reorthogonalization with previous (k=5) Lanczos vectors $\{q_1, \ldots, q_k\}$

(B)To (9.2.6): The duplicate pairs can occur!

$i \neq k$, $(\theta_i - \theta_k)\underbrace{y_i^T y_k}_{\text{O(1), if } y_i = y_k \Rightarrow Q_i \approx Q_k} = O(esp)$

How to avoid the duplicate pairs ?

The implicit Restart Lanczos algorithm:

Let $\{(\lambda_i, x_i)\}_{i=1}^n$: eigenpair of A

$u_1 = \sum_{i=1}^{k} r_i x_i + \sum_{i=k+1}^{n} r_i x_i$

$P(A)u_1 \leftarrow P(\lambda)$: Filter poly of degree $m-k$

$= \underbrace{\sum_{i=1}^{k} r_i P(\lambda_i) x_i}_{\text{expected}} + \underbrace{\sum_{i=k+1}^{n} r_i P(\lambda_i) x_i}_{\text{unexpected}}$

(1) $\lambda_{k+1}, \ldots, \lambda_m \in [a, b]$

$\lambda_1, \ldots, \lambda_k \notin [a, b]$

$P(\lambda)$ = Chebychev poly of degree m-k

(2) $\underbrace{u_1, \ldots, u_k}_{\text{expected}}, \underbrace{u_{k+1}, \ldots, u_m}_{\text{unexpected}}$: Ritz values

$P(t) = (t - \mu_{k+1}) \cdots (t - \mu_m)$

Implicit Restarted Algorithm:

$AQ_k = Q_k T_k + \underbrace{\beta_k q_{k+1}}_{r_k} e_k^T$, $k < m$

Lanczos $\Rightarrow AQ_m = Q_m T_m + \beta_m q_{m+1} e_m^T$

choose a filter poly of degree m-k

$P(t) = (t - \nu_1) \cdots (t - \nu_{m-k})$, $\nu_1, \ldots, \nu_{m-k}$: convergent Ritz values.

$\left(\begin{array}{c} \text{mathemetician:} P(A)u_1 = (A - \nu_1) \cdots (A - \nu_{m-k}) u_1 := q_1 \\ \text{Apply Lanczos on } q_1. \end{array} \right)$

$(A - K_1 I)Q_m = Q_m \underbrace{(T_m - K_1 I)}_{u_1 R_1 (QR - factorization)} + \beta_m q_{m+1} e_m^T$

$(A - K_1 I)\underbrace{Q_m u_1}_{Q_m{}^{(1)}} = (Q_m u_1)(R_1 u_1) + \beta_m q_{m+1}(e_m^T u_1)$

$AQ_m{}^{(1)} = Q_m{}^{(1)} T_m{}^{(1)} + \beta_m q_m b_{m+1}{}^{(1)T}$

where $b_{m+1}{}^{(1)T} = e_m^T u_1 \equiv (0, \ldots, 0, u_{m-1,m}{}^{(1)}, u_{m,m}{}^{(1)})$

$T_m{}^{(1)} = R_1 u_1 + K_1 I \equiv$ Tridiag

Remark: The first column of $Q_m{}^{(1)} e_1 = Q_m u_1 e_1$

$= \alpha(A - \nu_1 I)q_1 \equiv q_1{}^{(1)}$

Repeat this process with $\nu_2, \ldots, \nu_{m-k}$

$\Longrightarrow AQ_m{}^{(m-k)} = Q_m{}^{(m-k)} T_m{}^{(m-k)} + \beta_m q_{m+1} b_{m+1}{}^{(m-k)T}$

$Q_m{}^{(m-k)} e_1 := q_1 = (A - K_1 I) \cdots (A - K_{m-k} I) q_1$

Tumcate:

$AQ_k^{(m-k)} = Q_k^{(m-k)} T_{kk}^{(m-k)} + t_{k+1,k} q_{k+1}^{(m-k)} e_k^T + \beta_k u_{m,k} q_{m+1} e_k^T$

Let $\widetilde{Q}_k = Q_k^{(m-k)}$, $\widetilde{T}_{kk} = T_{kk}^{(m-k)}$,

$\widetilde{\beta}_k = ||t_{k+1,k}q_{k+1}^{(m-k)} + \beta_k u_{m,k} q_{m+1}||_2$,
$\widetilde{q}_{k+1} = \frac{\widetilde{P}_k}{\widetilde{\beta}_k}$, where $\widetilde{P}_k = t_{k+1,k}q_{k+1}^{(m-k)} + \beta_k u_{m,k} q_{m+1}$,
$A\widetilde{Q}_k = \widetilde{Q}_k = \widetilde{Q}_k \widetilde{T}_{kk} + \widetilde{\beta}_k \widetilde{q}_{k+1} e_k^T$.

(1) Implicit Restarted Lanczos
(2) Krylov-Schur cycle

[2] The symmetric eigenvalue problem , Parlett(1981)
CH.11 Approximation from a subspace
CH.12 Krylov subspace

Assumption: A: symmetric, $Az_i = \alpha_i z_i, i = 1, \ldots, n$.
$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$
$\alpha_{-n} \leq \cdots \leq \alpha_{-1}$
$\rho(x) = \rho(x, A) = \frac{x^T A x}{x^T x}$

Given a subspace $S^{(m)} =< F >=< F(F^TF)^{-\frac{1}{2}} >\equiv< Q >$
Rayleigh-Ritz-Quotient procedure

$H := \rho(Q) = Q^T A Q, Q^T Q = I$
$Hg_i = \theta_i g_i, (\theta_i g_i)$: Ritz pair
$y_i = Q_i g_i$(extension by Q)
Check : $\{(\theta_i, y_i)\}_{i=1}^m$approximate eigenpair?
$\|Ay_i - \theta_i y_i\|_2 \leq Tol, r_i = Ay_i - \theta_i y_i$residual

Optimality
(1) minmax:
$\alpha_j = \lambda_j(A) = \min\limits_{F^j \subseteq \mathbb{R}^n} \max\limits_{f \in F^j} \rho(f, A)$
$\beta_j := \min\limits_{G^j \subseteq S^m} \max\limits_{f \in G^j} \rho(f, A)$, $j \leq m$
$= \theta_j \equiv \lambda_j(H)$
$\because G^j \in S^m \Leftrightarrow Q\widetilde{G}^j = G^j$
$\therefore \beta_j := \min\limits_{\widetilde{G}^j \subseteq \mathbb{R}^m} \max\limits_{S \in \widetilde{G}^j} \rho(S.H) = \theta_j = \lambda_j(H)$, $j = 1, \ldots, m$.
(2) Optimal Residual:
Let $R(B) = AQ - QB$
$\Rightarrow \|R(H)\|_2 \leq \|R(B)\|_2$
$\|AQ - QH\|_2 \leq \|AQ - QB\|_2$
(3) Projection on S^m:
$\because Hg_i = \theta_i g_i$, $i = 1, \ldots, m$
$Q^T A Q g_i = \theta_i g_i$
$QQ^T A(Qg_i) = \theta_i (Qg_i)$
$Qg_i = y_i$, $QQ^T y_i = Q(Q^TQ)g_i = y_i$
$P_Q = QQ^T$ projection on $< Q >$
$(QQ^T)A(QQ^T)y_i = \theta_i y_i$
$(P_A A P_A)y_i = \theta_i y_i \Leftrightarrow P_A(Ay_i - \theta_i y_i) = 0$
$\Rightarrow r_i = Ay_i - \theta_i y_i \perp < Q >= S^m$

Theorem 7.1.9 $H = Q^T A Q, \exists \alpha_{\widetilde{i}} \in \sigma(A)$,
$s.t. |\theta_j - \alpha_{\widetilde{j}}| \leq \|R\|_2 = \|AQ - QH\|_2$

(By extension Thm)

Theorem 7.1.10 $\sum_{i=1}^{m}(\theta_j-\alpha_{\tilde{j}})^2 \le 2\|R\|_F^2$ *some* $\alpha_{\tilde{j}}$
Wiedlanclt - Hoffmann

Theorem 7.1.11 *Let y be a unit vector $\theta=\varrho(y)$, α be an EW of A closed to θ, z be the EV. Let* $r=\min_{\alpha_i\neq\alpha}|\lambda_i(A)-\theta|$.
Then (1)$|\theta-\alpha| \le \|r(y)\|^2/r$.
(2)$|\sin\psi| \le \|r(y)\|/r$,
where $r(y)=Ay-\theta y$, $\psi=\angle(y,z)$.

Proof: Claim(2): Decompose $y=z\cos\psi+w\sin\psi, z^Tw=0$.
$r(y)=z(\alpha-\theta)\cos\psi+(A-\theta)w\sin\psi$.
$\because Az=\alpha z \Rightarrow z^T(A-\theta)w=0$
$\Rightarrow \|r(y)\|_2^2=(\alpha-\theta)^2\cos^2\psi+\|(A-\theta)w\|_2^2\sin^2\psi$,
and $|w^T(A-\theta)(A-\theta)w|=\sum_{\alpha_i\neq\alpha}(\alpha_i-\theta)^2\xi_i^2, w=\sum_{\alpha_i\neq\alpha}\xi_i z_i$
$\ge r_2(\sum_{\alpha_i\neq\alpha}\xi_i^2)=r^2$,
$\Rightarrow \|r(y)\|_2^2 \ge \|(A-\theta)w\|_2^2\sin^2\psi$
$\Rightarrow |\sin\psi| \le \frac{\|r(y)\|_2}{r}$.

Claim(1): $r(y)\perp y$ $(r_i=Ay_i-\theta_i y_i \perp < Q >)$
ie. $0=y^Tr(y)=(\alpha-\theta)\cos^2\psi+w^T(A-\theta)w\sin^2\psi$
Thus $\frac{\cos^2\psi}{\sin^2\psi}=\frac{w^T(A-\theta)w}{\theta-\alpha}$

$$\left(\begin{array}{l}\text{Let}\frac{\cos^2\psi}{\sin^2\psi}\equiv k=\frac{w^T(A-\theta)w}{\theta-\alpha}\\ \Rightarrow \sin^2\psi\equiv\frac{1}{k+1}=\frac{\alpha-\theta}{w^T(A-\alpha)w}\\ \text{similarly},\cos^2\psi=\frac{k}{k+1}=\frac{w^T(A-\theta)w}{w^T(A-\alpha)w}\end{array}\right)$$

$\Rightarrow \|r(y)\|_2^2=(\theta-\alpha)w^T(A-\alpha)(A-\theta)w/w^T(A-\alpha)w$
$\because (A-\alpha)(A-\theta)z_i=(\alpha_i-\alpha)(\alpha_i-\theta)z_i$
positive definite >0
$w^T(A-\alpha)(A-\theta)w=\sum_{\alpha_i\neq\alpha}|\alpha_i-\alpha||\alpha_i-\theta|z_i^2$
$\ge r\sum_{\alpha_i\neq\alpha}|\alpha_i-\alpha|z_i^2 \ge r|\sum_{\alpha_i\neq\alpha}(\alpha_i-\alpha)z_i^2| = rw^T(A-\alpha)w$
$\Rightarrow |\theta-\alpha| \le \frac{\|r(y)\|_2^2}{r}$.

100 years old and still alive : Eigenvalue problems
Hank / G. Gloub / Van der Vorst / 2000
A priori bounds for interior Ritzvalues
Given $S^m =< Q >$ subspace, $\{(\theta_i,y_i)\}_{i=1}^m$ Ritz pairs of $H=Q^TAQ$
$Az_i=\alpha_i z_i$, $i=1,\ldots,n$.

Lemma 7.1.12 *For each $j\le m$ for any unit $s\in S^m$ satisfying $s^Tz_i=0$, $i=1,\ldots,j-1$. Then* $\alpha_j \le \theta_j \le \rho(s)+\sum_{i=1}^{j-1}(\alpha_{-1}-\theta_i)\sin^2\psi_i$

$$\leq \rho(s) + \sum_{i=1}^{j-1} (\alpha_{-1} - \alpha_i) \sin^2 \psi_i,$$

where $\psi_i = \angle(y_i, z_i)$.

Proof: Take $s = t + \sum_{i=1}^{j-1} r_i y_i$,

$t = \sum_{i=j}^{m} r_i y_i$, $t \perp y_i$, $i = 1, \ldots, j-1$.

Assumption: $s^T z_i = 0$, $i = 1, \ldots, j-1$.

Find bound of r_i, $i = 1, \ldots, j-1$.

$|r_i| = |s^T y_i| = |s^T (y_i - z_i \cos\psi)| \leq \|s\|_2 |\sin\psi|$.

$$\left(\begin{array}{rcl} \|y_i - z_i \cos\psi_i\|_2^2 & = & (y_i - z_i \cos\psi_i)^T (y_i - z_i \cos\psi_i) \\ & = & 1 - \cos^2\psi_i - \cos^2\psi_i + \cos^2\psi_i \\ & = & 1 - \cos^2\psi_i = \sin^2\psi_i \end{array} \right)$$

$\because t^T A y_i = 0$, and $y_i^T A y_k = 0$, $i \neq k$, $i = 1, \ldots, j-1$.

$\left(\; 0 = g_i^T (Q^T A Q) g_k = y_i^T A y_k, \; i \neq k, \text{ i,k=1,...,m.} \; \right)$

$\Rightarrow \rho(s) = t^T A t + \sum_{i=1}^{j-1} (y_i^T A y_i) r_i^2$

$$\begin{aligned} \rho(s) - \alpha_{-1} &= t^T (A - \alpha_{-1}) t + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1}) r_i^2 \\ &\geq \tfrac{t^T (A - \alpha_{-1}) t}{t^T t} + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1}) r_i^2 \\ &\geq \rho(t) - \alpha_{-1} + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1}) \sin^2 \varphi_i \end{aligned}$$

and $\rho(t) \geq \theta_j$, $t \perp y_i, i = 1, \ldots, j-1$

$\Rightarrow$ Assortion !

Let $\varphi_{ij} = \angle(z_i, y_j), i = 1, \ldots, n$, $j = 1, \ldots, m$,$\varphi_{ii} = \varphi_i$

$$y_j = \sum_{i=1}^{n} z_i \cos\varphi_{ij} \tag{1.22}$$

$$|\cos\varphi_{ij}| \leq |\sin\varphi_i| \tag{1.23}$$

$$\sum_{i=j+1}^{n} \cos^2\varphi_{ij} = \sin^2\varphi_j - \sum_{i=1}^{j-1} \cos^2\varphi_{ij} \tag{1.24}$$

Lemma 7.1.13 *For each* $j = 1, \ldots, m$,

$$\sin\varphi_j \leq [(\theta_j - \alpha_j) + \sum_{i=1}^{j-1} (\alpha_{j+1} - \alpha_i) \sin^2\varphi_i] / (\alpha_{j+1} - \alpha_j) \tag{1.25}$$

Remark:

prove (9.3.10):$|\cos\varphi_{ij}| = |y_j^T z_i| = |y_j^T (y_i \cos\varphi_i - z_i)| \quad (\because y_j^T y_i = 0, i \neq j)$

$$\leq \|y_j\|_2 \|y_i \cos\varphi_i - z_i\|_2 \leq |\sin\varphi_i|$$

$$(\because |(y_i \cos\varphi_i - z_i)^T (y_i \cos\varphi_i - z_i)| = \sin^2\varphi_i)$$

claim (1.24): $y_j = \sum_{i=1}^{n} z_i \cos\varphi_{ij}$

$$1 = (y_j, y_j) = \sum_{i=1}^{n} \cos^2 \varphi_{ij}$$

$$1 - \cos^2 \varphi_{jj} = \sin^2 \varphi_j = \sum_{i=j+1}^{n} \cos^2 \varphi_{ij} + \sum_{i=1}^{j-1} \cos^2 \varphi_{ij} \tag{1.26}$$

Proof: By (9.3.9), $\rho(y_j, A - \alpha_j I) = \theta_j - \alpha_j = \sum_{i=1}^{h} (\alpha_i - \alpha_j) \cos^2 \varphi_{ij}$

$$\theta_j - \alpha_j + \sum_{i=1}^{j-1} (\alpha_j - \alpha_i) \cos^2 \varphi_{ij} = \sum_{i=j+1}^{n} (\alpha_i - \alpha_j) \cos^2 \varphi_{ij}$$

$$\geq (\alpha_{j+1} - \alpha_j) \sum_{i=j+1}^{n} \cos^2 \varphi_{ij}$$

$$\overset{(9.3.13)}{=} (\alpha_{j+1} - \alpha_j)(\sin^2 \varphi_j - \sum_{i=1}^{j-1} \cos^2 \varphi_{ij})$$

Solve $\sin^2 \varphi_j$ and use (9.3.10)$\Rightarrow$ Inequation (9.3.12)

explanation: A priori bound for interior Ritz values

By Lemma 7.1.12, 7.1.13

$j = 1$: $\theta_1 \leq \rho(s), s^T z_1 = 0$(Lemma 7.1.12)

$j = 1$: $\sin^2 \varphi_1 \leq \frac{\theta_1 - \alpha_1}{\alpha_2 - \alpha_1} \leq \frac{\rho(s) - \alpha_1}{\alpha_2 - \alpha_1}, s^T z_1 = 0$(Lemma 7.1.13)

$j = 2$: $\theta_2 \leq \rho(s) + (\alpha_{-1} - \alpha_1) \sin^2 \varphi_1$

$\leq \rho(s) + (\alpha_{-1} - \alpha_1) \frac{\rho(\xi) - \alpha_1}{\alpha_2 - \alpha_1}$

$s^T z_1 = s^T z_2 = 0, \xi^T z_1 = 0$(Lemma 7.1.12)

$j = 2$: $\sin^2 \varphi_2 \overset{\text{(Lemma7.1.13)}}{\leq} (\theta_2 - \alpha_2) + \frac{(\alpha_3 - \alpha_1) \sin^2 \varphi_1}{\alpha_3 - \alpha_2}$

$\overset{j=1,j=2}{\leq} [\rho(s) + (\alpha_{-1} - \alpha_1)(\frac{\rho(t) - \alpha_1}{\alpha_2 - \alpha_1}) - \alpha_2] + \frac{\alpha_3 - \alpha_1}{\alpha_3 - \alpha_2}(\frac{\rho(t) - \alpha_1}{\alpha_2 - \alpha_1})$

$\vdots$

Chapter 12 Krylov subspace

$Az_j = \alpha_j z_j, j = 1, 2, \ldots, n$

$K^m(f) = [f, Af, \ldots, A^{m-1} f]$

$S_m = \mathcal{K}^m(f) = \langle f, Af, \ldots, A^{m-1} f \rangle$

created by Lanczos(A:symmetric) or Arnoldi(A:unsymmetric)

$S_m = \langle Q_m \rangle$

$H_m = (Q_m^T A Q_m), H_m S_j = \theta_j S_j, y_j = Q_m S_j, j = 1, \ldots, m$

(θ_j, y_j) : Rayleigh - Ritz pair , θ_j:R-Ritz value , y_j: R-Ritz vector

Lemma 7.1.14 *Let $\{(\theta_i, y_i)\}_{i=1}^m$ be Ritz pairs of $K^m(f)$, then $\omega(A)f \perp y_k \Leftrightarrow \omega(\theta_k) = 0, k = 1, \ldots, m$, where $\omega \in P^{m-1}$*

Proof: " $\Leftarrow$ " Let $\omega(\xi) = (\xi - \theta_k)\pi(\xi), \pi(\xi) \in P^{m-2}$

$y_k^T \omega(A) f = y_k^T (A - \theta_k)\pi(A) f \in K^m(f)$

$= r_k^T \pi(A) f$

$= 0$ ($\because r_k \perp \langle Q \rangle = \mathcal{K}^m(f)$)

" $\Rightarrow$ " exercise!

[2] The symmetric eigenvalue problem , Parlett(1981)

Definition 7.1.15 $\mu(\xi) = \prod_{i=1}^{m} (\xi - \theta_i), \pi_k(\xi) = \frac{\mu(\xi)}{(\xi - \theta_k)}$.

Corollary 7.1.16 $y_k = \frac{\pi_k(A)f}{\|\pi_k(A)f\|}$

Proof: $\because \pi_k(\theta_i) = 0, \theta_i \neq \theta_k, i \neq k$
$\overset{\text{Lemma7.1.14}}{\Rightarrow} \pi_k(A)f \perp y_i, \forall i \neq k$
$\Rightarrow \| \pi_k(A)f \| y_k \Rightarrow y_k = \frac{\pi_k(A)f}{\|\pi_k(A)f\|}$. ■

Lemma 7.1.17 *Let H be the normalized projection of f orthogonal to Z^j , $Z^j \equiv span(z_1, \ldots, z_j)$. For each $\pi \in P^{m-1}, j \leq m$,*

$$\rho(\pi(A)f, A - \alpha_j I) \leq (\alpha_n - \alpha_j)[\frac{\sin \angle(f, Z^j)}{\cos \angle(f, Z^j)} \frac{\|\pi(A)h\|}{|\pi(\alpha_j|)}]^2 \tag{1.27}$$

Proof: $\psi = \angle(f, Z^j) (= \cos^{-1} \|f^* Z^j\|)$
$f = g \cos\psi + h \sin\psi, \quad \because Z^j$is invariant.
$S \equiv \pi(A)f = \underbrace{\pi(A)g}_{\in Z^j} \cos\psi + \underbrace{\pi(A)h}_{\in (Z^j)^\perp} \sin\psi$
$\rho(s, A - \alpha_j I) = \frac{g^*(A - \alpha_j I)\pi^2(A)g \cos^2\psi + h^*(A - \alpha_j I)\pi^2(A)h \sin^2\psi}{\|\pi(A)f\|^2}$
$\because \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$
(a) $v^*(A - \alpha_j I)v \leq 0, \forall v \in Z^j$, in particular, $v = \pi(A)g$
$(\frac{v^*Av}{v^*v} \leq \alpha_j, \forall v \in Z^j)$
(b) $w^*(A - \alpha_j I)w \leq (\alpha_n - \alpha_j)\|w\|^2, \forall w \in (Z^j)^\perp$, in particular, $w = \pi(A)h$
reduction by $(a), (b) \Rightarrow \rho(s, A - \alpha_j I) \leq (\alpha_n - \alpha_j)[\frac{\|\pi(A)h\| \sin\psi}{\|\pi(A)f\|}]^2$
and $\|s\|^2 = \|\pi(A)f\|^2 \geq \pi^2(\alpha_j) \cos^2 \angle(f, z_j)$, where $f = \sum_{i=1}^{n} (f^* z_i) z_i$
$\Rightarrow$ (9.2.4). ■

The Error Bound of Kaniel and Saad:
The Error bounds come from choosing $\pi \in P^{m-1}$ in Lemma 7.1.17
s.t. (i) $|\pi(\alpha_j)|$ is large,while $\|\pi(A)h\|$ is small as possible,and
(ii) $\rho(s, A - \alpha_j I) \geq 0$
To (i): By Chebychev poly:

$$\|\pi(A)h\|^2 = \frac{\sum_{i=j+1}^{n} \pi^2(\alpha_i) \cos^2 \angle(f, z_j)}{\sum_{i=j+1}^{n} \cos^2 \angle(f, z_j)}$$
$$\leq \max_{i>j} \pi^2(\alpha_i) \leq \max_{\tau \in [\alpha_{j+1}, \alpha_n]} \pi^2(\tau)$$
Chebychev poly solves $\min_{\pi \in P^{n-j}} \max_{\tau \in [\alpha_{j+1}, \alpha_n]} \pi^2(\tau)$

To (ii): (a) $0 \le \theta_j - \alpha_j$ (Cauchy interlace Theorem)
(b) $\theta_j - \alpha_j \le \rho(s, A - \alpha_j I), s \perp y_i, i = 1, \ldots, j-1$
(By minmax theorem)
(c) $\theta_j - \alpha_j \le \rho(s, A - \alpha_j I) + \sum_{i=1}^{j-1} (\alpha_n - \alpha_i) \sin^2 \angle(y_i, z_i)$
if $s \perp z_i, i = 1, \ldots, j-1$, (by Chap11 Lemma 7.1.12)

Theorem 7.1.18 (Saad) *Let $\theta_1 \le \cdots \le \theta_n$ be the Ritz values from $\mathcal{K}^m(f)$ (by Lanczos or Arnoldi)*

For $j = 1, \ldots, m, \quad 0 \le \theta_j - \alpha_j \le (\alpha_n - \alpha_j)[\dfrac{\sin \angle(f, Z^j) \prod_{k=1}^{j-1} (\frac{\theta_k - \alpha_n}{\theta_k - \alpha_j})}{\cos \angle(f, Z^j) T_{m-j}(1 + 2r)}]^2$

where $r = \dfrac{\alpha_j - \alpha_{j+1}}{\alpha_{j+1} - \alpha_n}$

and $\tan \angle(z_j, \mathcal{K}^m) \le \dfrac{\sin \angle(f, Z^j) \prod_{k=1}^{j-1} (\frac{\alpha_k - \alpha_n}{\alpha_k - \alpha_j})}{\cos \angle(f, Z^j) T_{m-j}(1 + 2r)}$

Proof: Apply Lemma 7.1.17, Lemma 7.1.14,
To ensure (b), require $s \perp y_i, i = 1, \ldots, j-1$
By Lemma 7.1.14, we construct

$$\pi(\xi) = (\xi - \theta_1) \cdots (\xi - \theta_{j-1}) \tilde{\pi}(\xi), \tilde{\pi} \in P^{m-j}$$

$\because \pi(\theta_i) = 0 \Leftrightarrow \pi(A) f \perp y_i, \forall i = 1, \ldots, j-1$
By Lemma 7.1.17 for this $\pi(\xi)$:

$$\begin{aligned}
\frac{\|\pi(A)h\|}{|\pi(\alpha_j)|} &\le \frac{\|(A - \theta_1) \cdots (A - \theta_{j-1})\| \|\tilde{\pi}(A)h\|}{|(\alpha_j - \theta_1) \cdots (\alpha_j - \theta_{j-1})| |\tilde{\pi}(\alpha_j)|}, \ h \perp Z^j \\
&\le \prod_{k=1}^{j-1} |\frac{\alpha_n - \alpha_k}{\alpha_n - \theta_k}| \max_{\tau} \frac{|\tilde{\pi}(\tau)|}{|\tilde{\pi}(\alpha_j)|}, \ \tau \in [\alpha_{j+1}, \alpha_j] \\
&\le \prod_{k=1}^{j-1} |\frac{\alpha_n - \alpha_k}{\alpha_j - \alpha_k}| \min_{\tilde{\pi} \in P^{m-j}} \max_{j} \frac{|\tilde{\pi}(\tau)|}{|\tilde{\pi}(\alpha_j)|} \\
&= \prod_{k=1}^{j-1} |\frac{\alpha_n - \alpha_k}{\alpha_j - \alpha_k}| \frac{1}{T_{m-j}(1 + 2r)} \qquad (1.28)
\end{aligned}$$

$$(t \in [\alpha_{j+1}, \alpha_n] \longrightarrow \tilde{t} \in [-1, 1], \tilde{t} = \frac{2t - \alpha_{j+1} - \alpha_n}{\alpha_n - \alpha_{j+1}})$$

$$0 \le \theta_j - \alpha_j \overset{(9.2.4),(9.2.5)}{\le} (\alpha_n - \alpha_j)[\frac{\sin \angle(f, Z^j) \prod_{k=1}^{j-1} (\frac{\theta_k - \alpha_m}{\theta_k - \alpha_j})}{\cos \angle(f, Z^j) T_{m-j}(1 + 2r)}]^2$$

To prove the second inequality:
π is chosen to satisfy $\pi(\alpha_i) = 0, i = 1, \ldots, j-1$
$s = \pi(A) f = z_j \pi(\alpha_j) \cos \angle(f, z_j) + \pi(A) h \sin \psi$

$\Rightarrow \tan\angle(s, z_j) = \dfrac{\sin\angle(f, Z^j)\|\pi(A)h\|}{\cos\angle(f, z_j)|\pi(\alpha_j)|}$,
where $\pi(\xi) = (\xi - \alpha_1)\cdots(\xi - \alpha_{j-1})\tilde{\pi}(\xi), \tilde{\pi}(\xi) \in P^{m-j}$
$\tilde{\pi}$ is chosen by chebychev poly as above $\Rightarrow$ inequality. ■

Theorem 7.1.19 *Let* $\theta_{-m} \le \ldots \le \theta_{-1}$ *be Royleigh-Ritz values of* $\mathcal{K}^m(f)$, $Az_{-j} = \alpha_{-j}z_{-j}, j = n, \ldots, 1, \alpha_{-n} \le \cdots \le \alpha_{-1}$, $0 \le \alpha_{-j} - \theta_{-j} \le (\alpha_{-j} - \alpha_{-1})[\dfrac{\sin\angle(f, Z^{-j}) \prod_{k=-j+1}^{-1} (\frac{\alpha_{-n} - \theta_{-k}}{\alpha_{-k} - \theta_{-j}})}{\cos\angle(f, z_{-j})T_{m-j}(1 + 2r)}]^2$,

where $r = \dfrac{\alpha_{-j-1} - \alpha_{-j}}{\alpha_{-n} - \alpha_{-j-1}}$, $\tan(z_{-j}, \mathcal{K}^m) \le \dfrac{\sin\angle(f, Z^{-j})}{\cos\angle(f, z_{-j})}[\dfrac{\prod_{k=-j+1}^{-1} (\frac{\alpha_{-k} - \alpha_{-n}}{\alpha_{-k} - \alpha_{-j}})}{T_{m-j}(1 + 2r)}]^2$.

Theorem 7.1.20 (Kaniel)

By (c) and Lemma 7.1.12 of Chap11

$$s = \pi(A)f = (A - \alpha_1)\cdots(A - \alpha_{j-1})\tilde{\pi}(A)f$$

By Lemma 7.1.17 and Lemma 7.1.14

$$\Rightarrow 0 \le \theta_j - \alpha_j \le (\alpha_n - \alpha_j)[\frac{\sin\angle(f, Z^j) \prod_{k=1}^{j-1} (\frac{\alpha_k - \alpha_n}{\alpha_k - \alpha_j})}{\cos\angle(f, z_j)T_{m-j}(1 + 2r)}]^2 + \sum_{k=1}^{j-1} (\alpha_n - \alpha_k)\sin^2\angle(y_k, z_k)$$

$$\textit{and } \sin^2\angle(y_j, z_j) \le \frac{(\theta_j - \alpha_j) + \sum_{k=1}^{j-1} (\alpha_{j+1} - \alpha_k)\sin^2\angle(y_k, z_k)}{\alpha_{j+1} - \alpha_j}$$

where $r = \dfrac{\alpha_j - \alpha_{j+1}}{\alpha_{j+1} - \alpha_n}$.

7.2 Applications to linear Systems and Least Squares

7.2.1 Symmetric Positive Definite System

Recall: Let A be symmetric positive definite and $Ax^* = b$. Then x^* minimizes the functional

$$\phi(x) = \frac{1}{2}x^T Ax - b^T x. \tag{2.1}$$

An approximate minimizer of ϕ can be regarded as an approximate solution to $Ax = b$.

One way to produce a sequence $\{x_j\}$ that converges to x^* is to generate a sequence of orthonormal vectors $\{q_j\}$ and to let x_j minimize ϕ over $\text{span}\{q_1, \cdots, q_j\}$, where $j = 1, \cdots, n$. Let $Q_j = [q_1, \cdots, q_j]$. Since

$$x \in \text{span}\{q_1, \cdots, q_j\} \;\Rightarrow\; \phi(x) = \frac{1}{2}y^T(Q_j^T A Q_j)y - y^T(Q_j^T b)$$

for some $y \in R^j$, it follows that

$$x_j = Q_j y_j, \tag{2.2}$$

where

$$(Q_j^T A Q_j)y_j = Q_j^T b. \tag{2.3}$$

Note that $Ax_n = b$.

We now consider how this approach to solving $Ax = b$ can be made effective when A is large and sparse. There are two hurdles to overcome:

(1) the linear system (2.3) must be easily solved;

(2) we must be able to compute x_j without having to refer to $q_1, \cdots, q_j$ explicitly as (2.2) suggests.

To (1): we use Lanczos algorithm algorithm 7.1.1 to generate the q_i. After j steps we obtain

$$AQ_j = Q_j T_j + r_j e_j^T, \tag{2.4}$$

where

$$T_j = Q_j^T A Q_j = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ 0 & & \beta_{j-1} & \alpha_j \end{bmatrix} \text{ and } T_j y_j = Q_j^T b. \tag{2.5}$$

With this approach, (2.3) becomes a symmetric positive definite tridiagonal system which may be solved by LDL^T Cholesky decomposition, i.e.,

$$T_j = L_j D_j L_j^T, \tag{2.6}$$

where

$$L_j = \begin{bmatrix} 1 & & & 0 \\ \mu_2 & \ddots & & \vdots \\ & \ddots & \ddots & 0 \\ 0 & & \mu_j & 1 \end{bmatrix} \quad \text{and} \quad D_j = \begin{bmatrix} d_1 & & 0 \\ & \ddots & 0 \\ 0 & & d_j \end{bmatrix}.$$

Comparsion of the entries of (2.6):

$$\begin{array}{rcl} d_1 & = & \alpha_1, \\ i & = & 2, \cdots, j, \\ \mu_i & = & \beta_{i-1}/d_{i-1}, \\ d_i & = & \alpha_i - \beta_{i-1}\mu_i. \end{array} \tag{2.7}$$

Note that we need only calculate

$$\begin{array}{rcl} \mu_j & = & \beta_{j-1}/d_{j-1} \\ d_j & = & \alpha_j - \beta_{j-1}\mu_j \end{array} \tag{2.8}$$

in order to obtain L_j and D_j from L_{j-1} and D_{j-1}.

To (2): Trick: we define $C_j = [c_1, \cdots, c_j] \in R^{n\times j}$ and $p_j \in R^j$ by the equations

$$\begin{array}{rcl} C_j L_j^T & = & Q_j \\ L_j D_j p_j & = & Q_j^T b \end{array} \tag{2.9}$$

and observe that

$$x_j = Q_j T_j^{-1} Q_j^T b = Q_j (L_j D_j L_j^T)^{-1} Q_j^T b = C_j p_j.$$

It follows from (2.9) that

$$[c_1, \mu_2 c_1 + c_2, \cdots, \mu_j c_{j-1} + c_j] = [q_1, \cdots, q_j],$$

and therefore

$$C_j = [C_{j-1}, c_j], \quad c_j = q_j - \mu_j c_{j-1}.$$

If we set $p_j = [\rho_1, \cdots, \rho_j]^T$ in $L_j D_j p_j = Q_j^T b$, then that equation becomes

$$\left[\begin{array}{c|c} L_{j-1}D_{j-1} & 0 \\ \hline 0\cdots 0\,\mu_j d_{j-1} & d_j \end{array}\right] \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{j-1} \\ \rho_j \end{bmatrix} = \begin{bmatrix} q_1^T b \\ q_2^T b \\ \vdots \\ q_{j-1}^T b \\ q_j^T b \end{bmatrix}.$$

Since $L_{j-1}D_{j-1}p_{j-1} = Q_{j-1}^T b$, it follows that

$$p_j = \begin{bmatrix} p_{j-1} \\ \rho_j \end{bmatrix}, \quad \rho_j = (q_j^T b - \mu_j d_{j-1}\rho_{j-1})/d_j$$

and thus

$$x_j = C_j p_j = C_{j-1} p_{j-1} + \rho_j c_j = x_{j-1} + \rho_j c_j.$$

This is precisely the kind of recursive formula for x_j that we need. Together with (2.8) and (2.9) it enables us to make the transition from $(q_{j-1}, c_{j-1}, x_{j-1})$ to (q_j, c_j, x_j) with a minimal amount of work and storage.

A further simplification results if we set $q_1 = b/\beta_0$ where $\beta_0 = \|b\|_2$. For this choice of a Lanczos starting vector we see that $q_i^T b = 0$ for $i = 2, 3, \cdots$. It follows from (2.4) that

$$\begin{aligned} Ax_j = AQ_j y_j &= Q_j T_j y_j + r_j e_j^T y_j = Q_j Q_j^T b + r_j e_j^T y_j \\ &= b + r_j e_j^T y_j. \end{aligned}$$

Thus, if $\beta_j = \|r_j\|_2 = 0$ in the Lanczos iteration, then $Ax_j = b$. Moreover, since $\|Ax_j - b\|_2 = \beta_j |e_j^T y_j|$, the iteration provides estimates of the current residual.

Algorithm 7.2.1 Given $b \in R^n$ and a symmetric positive definite $A \in R^{n\times n}$. The following algorithm computes $x \in R^n$ such that $Ax = b$.

$\beta_0 = \|b\|_2, q_1 = b/\beta_0, \alpha_1 = q_1^T A q_1, d_1 = \alpha_1, c_1 = q_1, x_1 = b/\alpha_1.$
For $j = 1, \cdots, n-1$,
  $r_j = (A - \alpha_j) q_j - \beta_{j-1} q_{j-1} \quad (\beta_0 q_0 \equiv 0),$
  $\beta_j = \|r_j\|_2,$
  If $\beta_j = 0$ then
    Set $x^* = x_j$ and stop; else
    $q_{j+1} = r_j/\beta_j,$
    $\alpha_{j+1} = q_{j+1}^T A q_{j+1},$
    $\mu_{j+1} = \beta_j/d_j,$
    $d_{j+1} = \alpha_{j+1} - \mu_{j+1}\beta_j,$
    $\rho_{j+1} = -\mu_{j+1} d_j \rho_j / d_{j+1},$
    $c_{j+1} = q_{j+1} - \mu_{j+1} c_j,$
    $x_{j+1} = x_j + \rho_{j+1} c_{j+1},$
    $x^* = x_n.$

This algorithm requires one matrix-vector multiplication and $5n$ flops per iteration.

Symmetric Indefinite Systems

A key feature in the above development is the idea of computing LDL^T Cholesky decomposition of tridiagonal T_j. Unfortunately, this is potentially unstable if A, and consequently T_j, is not positive definite. Paige and Saunders (1975) had developed the recursion for x_j by an LQ decomposition of T_j. At the j-th step of the iteration we will Given rotations $J_1, \cdots, J_{j-1}$ such that

$$T_j J_1 \cdots J_{j-1} = L_j = \begin{bmatrix} d_1 & & & & 0 \\ e_2 & d_2 & & & \\ f_3 & e_3 & d_3 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & f_j & e_j & d_j \end{bmatrix}.$$

Note that with this factorization, x_j is given by

$$x_j = Q_j y_j = Q_j T_j^{-1} Q_j^T b = W_j s_j,$$

where $W_j \in R^{n\times j}$ and $s_j \in R^j$ are defined by

$$W_j = Q_j J_1 \cdots J_{j-1} \quad \text{and} \quad L_j s_j = Q_j^T b.$$

Scrutiny of these equations enables one to develop a formula for computing x_j from x_{j-1} and an easily computed multiple of w_j, the last column of W_j.

Connection of Algorithm 14.1 and CG method:
Let

x_j^L : Iterative vector generated by Algorithm 7.2.1

x_i^{CG} : Iterative vector generated by CG method with , $x_0^{CG} = 0$.

Since $r_0^{CG} = b - Ax_0 = b = p_0^{CG}$, then

$$x_1^{CG} = \alpha_0^{CG} p_0 = \frac{b^T b}{b^T A b} b = x_1^L.$$

Claim: $x_i^{CG} = x_i^L$ for $i = 1, 2, \cdots$,
(1) CG method (A variant version):

$$\begin{aligned}
&x_0 = 0, r_0 = b,\\
&\quad \text{For } k = 1, \cdots, n,\\
&\qquad \text{if } r_{k-1} = 0 \text{ then set } x = x_{k-1} \text{ and quit.}\\
&\qquad \text{else } \beta_k = r_{k-1}^T r_{k-1} / r_{k-2}^T r_{k-2} \quad (\beta_1 \equiv 0),\\
&\qquad\quad p_k = r_{k-1} + \beta_k p_{k-1} \quad (p_1 \equiv r_0),\\
&\qquad\quad \alpha_k = r_{k-1}^T r_{k-1} / p_k^T A p_k,\\
&\qquad\quad x_k = x_{k-1} + \alpha_k p_k,\\
&\qquad\quad r_k = r_{k-1} - \alpha_k A p_k,\\
&\quad x = x_n.
\end{aligned} \tag{2.10}$$

Define $R_k = [r_0, \cdots, r_{k-1}] \in R^{n\times k}$ and

$$B_k = \begin{bmatrix} 1 & -\beta_2 & & 0 \\ & 1 & \ddots & \\ & & \ddots & -\beta_k \\ 0 & & & 1 \end{bmatrix}.$$

From $p_j = r_{j-1} + \beta_j p_{j-1}$ $(j = 2, \cdots, k)$ and $p_1 = r_0$ it follows $R_k = P_k B_k$. Since the columns of $P_k = [p_1, \cdots, p_k]$ are A-conjugate, we see that

$$R_k^T A R_k = B_k^T diag(p_1^T A p_1, \cdots, p_k^T A p_k) B_k$$

is tridiagonal. Since span$\{p_1, \cdots, p_k\}$=span$\{r_0, \cdots, r_{k-1}\}$=span$\{b, Ab, \cdots, A^{k-1}b\}$ and $r_0, \cdots, r_{k-1}$ are mutually orthogonal, it follows that if

$$\triangle_k = diag(\beta_0, \cdots, \beta_{k-1}), \qquad \beta_i = \|r_i\|_2,$$

then the columns of $R_k \triangle_k^{-1}$ form an orthonormal basis for span$\{b, Ab, \cdots, A^{k-1}b\}$. Consequently the columns of this matrix are essentially the Lanczos vectors of algorithm 7.2.1, i.e., $q_i^L = \pm r_{i-1}^{CG}/\beta_{i-1}$ $(i = 1, \cdots, k)$. Moreover,

$$T_k = \triangle_k^{-1} B_k^T diag(p_i^T A p_i) B_k \triangle_k^{-1} .$$

The diagonal and subdiagonal of this matrix involve quantities that are readily available during the conjugate gradient iteration. Thus, we can obtain good estimate of $A's$ extremal eigenvalues (and condition number) as we generate the x_k in (2.11).

$$p_i^{CG} = c_i^L \cdot \text{constant}.$$

Show that: c_i^L are A-orthogonal.

$$C_j L_j^T = Q_j \Rightarrow C_j = Q_j L_j^{-T} \Rightarrow$$

$$\begin{aligned} C_j^T A C_j &= L_j^{-1} Q_j^T A Q_j L_j^{-T} = L_j^{-1} T_j L_j^{-T} \\ &= L_j^{-1} L_j D_j L_j^T L_j^{-T} = D_j. \end{aligned}$$

So $\{c_i\}_{i=1}^j$ are A-orthogonal.

(2) It is well known that x_j^{CG} minimizes the functional $\phi(x) = \frac{1}{2}x^T Ax - b^T x$ in the subspace span$\{r_0, Ar_0, \cdots, A^{j-1}r_0\}$ and x_j^L minimize $\phi(x) = \frac{1}{2}x^T Ax - b^T x$ in the subspace span$\{q_1, \cdots, q_j\}$. We also know that $K[q_1, A, j] = Q_j R_j$ which implies $\mathcal{K}(q_1, A, j) =$span $\{q_1, \cdots, q_j\}$. But $q_1 = b/\|b\|_2$, $r_0 = b$, so span $\{r_0, Ar_0, \cdots, A^{j-1}r_0\} = \mathcal{K}(q_1, A, j) =$span $\{q_1, \cdots, q_j\}$ therefore we have $x_j^{CG} = x_j^L$.

7.2.2 Bidiagonalization and the SVD

Suppose $U^T AV = B$ the bidiagonalization of $A \in R^{m\times n}$ and that

$$\begin{aligned} U &= [u_1, \cdots, u_m], \quad U^T U = I_m, \\ V &= [v_1, \cdots, v_n], \quad V^T V = I_n, \end{aligned} \tag{2.11}$$

and

$$B = \left[\begin{array}{cccc} \alpha_1 & \beta_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \beta_{n-1} \\ 0 & & & \alpha_n \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right]. \tag{2.12}$$

Recall that this decomposition serves as a front end for the SVD algorithm. Unfortunately, if A is large and sparse, then we can expect large, dense submatrices to arise during the Householder transformation for the bidiagonalization. It would be nice to develop a method for computing B directly without any orthogonal update of the matrix A.

We compare columns in the equations $AV = UB$ and $A^T U = VB^T$:

$$Av_j = \alpha_j u_j + \beta_{j-1} u_{j-1}, \quad \beta_0 u_0 \equiv 0, A^T u_j = \alpha_j v_j + \beta_j v_{j+1}, \quad \beta_n v_{n+1} \equiv 0,$$

for $j = 1, \cdots, n$. Define

$$r_j = Av_j - \beta_{j-1}u_{j-1} \text{ and } p_j = A^T u_j - \alpha_j v_j.$$

We may conclude that

$$\begin{aligned} \alpha_j &= \pm\|r_j\|_2, \quad u_j = r_j/\alpha_j, \\ v_{j+1} &= p_j/\beta_j, \quad \beta_j = \pm\|p_j\|_2. \end{aligned}$$

These equations define the Lanczos method for bidiagonaling a rectangular matrix (by Paige (1974)):

$$\begin{array}{l} \text{Given } v_1 \in R^n, \text{ with unit 2-norm.} \\ r_1 = Av_1, \ \ \alpha_1 = \|r_1\|_2. \\ \text{For } j = 1, \cdots, n, \\ \text{If } \alpha_j = 0 \text{ then stop; else} \\ u_j = r_j/\alpha_j, \ p_j = A^T u_j - \alpha_j v_j, \ \beta_j = \|p_j\|_2, \\ \text{If } \beta_j = 0 \ \text{ then stop; else} \\ v_{j+1} = p_j/\beta_j, r_{j+1} = Av_{j+1} - \beta_j u_j, \alpha_{j+1} = \|r_{j+1}\|_2. \end{array} \tag{2.13}$$

It is essentially equivalent to applying the Lanczos tridiagonalization scheme to the symmetric matrix $C = \begin{bmatrix} O & A \\ A^T & 0 \end{bmatrix}$. We know that

$$\lambda_i(C) = \sigma_i(A) = -\lambda_{n+m-i+1}(C)$$

for $i = 1, \cdots, n$. Because of this, the large singular values of the bidiagonal matrix $B_j = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \beta_{j-1} \\ 0 & & & \alpha_j \end{bmatrix}$ tend to be very good approximations to the large singular values of A.

Least Squares: As detailed in chapter III the full-rank LS problem $\min\|Ax - b\|_2$ can be solved by the bidiagonalization (2.11)-(2.12). In particular,

$$x_{LS} = Vy_{LS} = \sum_{i=1}^{n} a_i v_i,$$

where $y = (a_1, \cdots, a_n)^T$ solves the bidiagonal system $By = (u_1^T b, \cdots, u_n^T b)^T$.

Disadvantage: Note that because B is upper bidiagonal, we cannot solve for y until the bidiagonalization is complete. We are required to save the vectors $v_1, \cdots, v_n$ an unhappy circumstance if n is very large.

Modification: It can be accomplished more favorably if A is reduced to lower bidiagonal

form:

$$U^TAV = B = \left[\begin{array}{cccc} \alpha_1 & & & 0 \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_n \\ 0 & & & \beta_n \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right], \qquad m \geq n+1,$$

where $V = [v_1, \cdots, v_n]$ and $U = [u_1, \cdots, u_m]$. It is straightforward to develop a Lanczos procedure which is very similar to (2.13). Let $V_j = [v_1, \cdots, v_j]$, $U_j = [u_1, \cdots, u_j]$ and

$$B_j = \left[\begin{array}{cccc} \alpha_1 & & & 0 \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_j \\ 0 & & & \beta_j \end{array}\right] \in R^{(j+1)\times j}$$

and consider minimizing $\|Ax - b\|_2$ over all vectors of the form $x = V_j y$, $y \in R^j$. Since

$$\|AV_jy - b\|_2 = \|U^TAV_jy - U^Tb\|_2 = \|B_jy - U_{j+1}^Tb\|_2 + \sum_{i=j+2}^{m}(u_i^Tb)^2,$$

it follows that $x_j = V_jy_j$ is the minimizer of the LS problem over span$\{V_j\}$, where y_j minimizes the $(j+1)\times j$ LS problem $\min\|B_jy - U_{j+1}^Tb\|_2$. Since B_j is lower bidiagonal, it is easy to compute Jacobi rotations $J_1, \cdots, J_j$ such that

$$J_j \cdots J_1 B_j = \left[\begin{array}{c} R_j \\ 0 \end{array}\right] \text{ is upper bidiagonal.}$$

Let $J_j \cdots J_1 U_{j+1}^T b = \left[\begin{array}{c} d_j \\ u \end{array}\right]$, then

$$\|B_jy - U_{j+1}^Tb\|_2 = \|J_j \cdots J_1 y - J_j \cdots J_1 U_{j+1}^T b\|_2 = \| \left[\begin{array}{c} R_j \\ 0 \end{array}\right] y - \left[\begin{array}{c} d_j \\ u \end{array}\right] \|_2.$$

So $y_j = R_j^{-1}d_j$, $x_j = V_jy_j = V_jR_j^{-1}d_j = W_jd_j$. Let

$$W_j = (W_{j-1}, w_j), w_j = (v_j - w_{j-1}r_{j-1,j})/r_{jj}$$

($r_{j-1,j}$ and r_{jj} are elements of R_j). R_j can be computed from R_{j-1}. Similarly, $d_j = \left[\begin{array}{c} d_{j-1} \\ \delta_j \end{array}\right]$, x_j can be obtained from x_{j-1}:

$$x_j = W_jd_j = (W_{j-1}, w_j) \left[\begin{array}{c} d_{j-1} \\ \delta_j \end{array}\right] = W_{j-1}d_{j-1} + w_j\delta_j.$$

Thus

$$x_j = x_{j-1} + w_j\delta_j.$$

For details see Paige-Saunders (1978).

Error Estimation of LS-problem

Continuity of A^+ of the function: $R^{m\times n} \to R^{m\times n}$ defined by $A \longmapsto A^+$.

Lemma 7.2.2 *If $\{A_i\}$ converges to A and $rank(A_i) = rank(A) = n$, then $\{A_i^+\}$ also converges to A^+.*

Proof: Since $\lim\limits_{i\to\infty} A_i^T A_i = A^T A$ nonsingular, so

$$A_i^+ = (A_i^T A_i)^{-1} A_i^T \overset{i\to\infty}{\longrightarrow} (A^T A)^{-1} A^T = A^+.$$

■

Example: Let $A_\epsilon = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \\ 0 & 0 \end{bmatrix}$, $\epsilon > 0$, $A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ then $A_\epsilon \to A_0$ as $\epsilon \to 0$, $rank(A_0) < 2$. But $A_\epsilon^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\epsilon & 0 \end{bmatrix} \not\longrightarrow A_0^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ as $\epsilon \to 0$.

Theorem 7.2.3 *Let $A, B \in R^{m\times n}$, then holds*

$$\|A^+ - B^+\|_F \le \sqrt{2}\|A - B\|_F \max\{\|A^+\|_2^2, \|B^+\|_2^2\}.$$

Without proof.

Remark: It does not follow that $A \to B$ implies $A^+ \to B^+$. Because A^+ can diverges to ∞, see example.

Theorem 7.2.4 *If $rank(A) = rank(B)$ then*

$$\|A^+ - B^+\|_F \le \mu\|A^+\|_2\|B^+\|_2\|A - B\|_F,$$

where

$$\mu = \begin{cases} \sqrt{2}, & \text{if } rank(A) < \min(m,n) \\ 1, & \text{if } rank(A) = \min(m,n). \end{cases}$$

Pseudo-Inverse of A: A^+ is the unique solution of equations

$$A^+AA^+ = A^+, \quad (AA^+)^* = AA^+,$$
$$AA^+A = A, \quad (A^+A)^* = A^+A.$$

$P_A = AA^+$ is Hermitian. P_A is idempotent, and $R(P_A) = R(A)$. P_A is the orthogonal projection onto $R(A)$. Similarly, $R(A) = A^+A$ is the projection onto $R(A^*)$. Furthermore,

$$\rho_{LS}^2 = \|b - AA^+b\|_2^2 = \|(I - AA^+)b\|_2^2.$$

Banach Lemma: $\|B^{-1} - A^{-1}\| \le \|A - B\|\|A^{-1}\|\|B^{-1}\|$.
Proof: From $((A+\delta A)^{-1} - A^{-1})(A+\delta A) = I - I - A^{-1}\delta A$ follows Lemma immediately.
■

Theorem 7.2.5 *(1) The product $P_B P_A^{\perp}$ can be written in the form*

$$P_B P_A^{\perp} = (B^+)^* R_B E^* P_A^{\perp},$$

where $P_A^{\perp} = I - P_A$, $B = A + E$. Thus $\|P_B P_A^{\perp}\| \leq \|B^+\|_2 \|E\|$.

(2) If $rank(A) = rank(B)$, then $\|P_B P_A^{\perp}\| \leq \min\{\|B^+\|_2, \|A^+\|_2\}\|E\|$.

Proof:

$$\begin{aligned} P_B P_A^{\perp} &= P_B^* P_A^{\perp} = (B^+)^* B^* P_A^{\perp} = (B^+)^*(A+E)^* P_A^{\perp} = (B^+)^* E^* P_A^{\perp} \\ &= (B^+)^* B^* (B^+)^* E^* P_A^{\perp} = (B^+)^* R_B E^* P_A^{\perp} \quad (\|R_B\| \leq 1, \|P_A^{\perp}\| \leq 1). \end{aligned}$$

Part (2) follows from the fact that $rank(A) \leq rank(B) \Rightarrow \|P_B P_A^{\perp}\| \leq \|P_B^{\perp} P_A\|$. Exercise! (Using C-S decomposition). ■

Theorem 7.2.6 *It holds*

$$B^+ - A^+ = -\overbrace{B^+ P_B E R_A A^+}^{F_1} + \overbrace{B^+ P_B P_A^{\perp}}^{F_2} - \overbrace{R_B^{\perp} R_A A^+}^{F_3}.$$
$$B^+ - A^+ = -B^+ P_B E R_A A^+ + (B^* B)^+ R_B E^* P_A^{\perp} - R_B^{\perp} E^* P_A (AA^*)^+.$$

Proof:

$$\begin{aligned} & -B^+ B B^+ (B-A) A^+ A A^+ + B^+ B B^+ (I - AA^+) - (I - B^+ B)(A^+ A) A^+ \\ = \; & -B^+ (B-A) A^+ + B^+ (I - AA^+) - (I - B^+ B) A^+ \\ = \; & B^+ - A^+ \text{ (Substitute } P_B = BB^+, \, E = B - A, \, R_A = AA^+, \cdots .). \end{aligned}$$

■

Theorem 7.2.7 *If $B = A + E$, then*

$$\|B^+ - A^+\|_F \leq \sqrt{2}\|E\|_F \max\{\|A^+\|_2^2, \|B^+\|_2^2\}.$$

Proof: Suppose $rank(B) \leq rank(A)$. Then the column spaces of F_1 and F_2 are orthogonal to the column space of F_3. Hence

$$\|B^+ - A^+\|_F^2 = \|F_1 + F_2\|_F^2 + \|F_3\|_F^2 \quad ((I - B^+ B) B^+ = 0).$$

Since $F_1 + F_2 = B^+(P_B E A^+ P_A + P_B P_A^{\perp})$, we have

$$\|F_1 + F_2\|_F^2 \leq \|B^+\|_2^2 (\|P_B E A^+ P_A\|_F^2 + \|P_B P_A^{\perp}\|_F^2).$$

By theorem 7.2.5 and 7.2.6 follows that

$$\begin{aligned} & \|P_B E A^+ P_A\|_F^2 + \|P_B P_A^{\perp}\|_F^2 \leq \|P_B E A^+\|_F^2 + \|P_B^{\perp} P_A\|_F^2 \\ = \; & \|P_B E A^+\|_F^2 + \|P_B^{\perp} E A^+\|_F^2 = \|E A^+\|_F^2 \leq \|E\|_F^2 \|A^+\|_2^2. \end{aligned}$$

Thus

$$\|F_1 + F_2\|_F \leq \|A^+\|_2 \|B^+\|_2 \|E\|_F \quad (P_B^{\perp} P_A = P_B^{\perp} E R_A A^+ = P_B^{\perp} E A^+).$$

By theorem 7.2.6 we have

$$\begin{aligned} \|F_3\|_F &\leq \|A^+\|_2 \|R_B^{\perp} R_A\|_F = \|A^+\|_2 \|R_A R_B^{\perp}\|_F \\ &= \|A^+\|_2 \|A^+ E R_B^{\perp}\| \leq \|A^+\|_2^2 \|E\|_F. \end{aligned}$$

The final bound is symmetric in A and B, it also holds when $rank(B) \geq rank(A)$. ■

Theorem 7.2.8 *If $rank(A) = rank(B)$, then*

$$\|B^+ - A^+\|_F \leq \sqrt{2}\|A^+\|_2\|B^+\|_2\|E\|_F. \quad \textit{(see Wedin (1973))}$$

■

From above we have

$$\frac{\|B^+ - A^+\|_F}{\|B^+\|_2} \leq \sqrt{2}k_2(A)\frac{\|E\|_F}{\|A\|_2}.$$

This bound implies that as E approaches zero, the relative error in B^+ approaches zero, which further implies that B^+ approach A^+.

Corollary 7.2.9 $\lim_{B\to A} B^+ = A^+ \Longleftrightarrow rank(A) = rank(B)$ *as B appraches A.*

(See Stewart 1977) ■

Perturbation of solutions of the LS-problem

We first state two Corollarys of Theorem (SVD).

Theorem 7.2.10 (SVD) *If $A \in R^{m\times n}$ then there exists orthogonal matrices $U = [u_1, \cdots, u_m] \in R^{m\times m}$ and $V = [v_1, \cdots, v_n] \in R^{n\times n}$ such that $U^TAV = diag(\sigma_1, \cdots, \sigma_p)$, $p = min(m, n)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.*

Corollary 7.2.11 *If the SVD is given by theorem 7.2.10 and $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, then*
(1) $rank(A) = r$.
(2) $\mathcal{N}(A) = span\{v_{r+1}, \cdots, v_n\}$.
(3) $Range(A) = span\{u_1, \cdots, u_r\}$.
(4) $A = \Sigma_{i=1}^r \sigma_i u_i v_i^T = U_r\Sigma_r V_r^T$, where $U_r = [u_1, \cdots, u_r]$, $V_r = [v_1, \cdots, v_r]$ and $\Sigma_r = diag(\sigma_1, \cdots, \sigma_r)$.
(5) $\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$.
(6) $\|A\|_2 = \sigma_1$.

Proof: exercise !

Corollary 7.2.12 *Let SVD of $A \in R^{m\times n}$ is given by theorem 7.2.10. If $k < r = rank(A)$ and $A_k = \Sigma_{i=1}^k \sigma_i u_i v_i^T$, then*

$$\min_{rank(X)=k, X\in R^{m\times n}} \|A - X\|_2 = \|A - A_k\|_2 = \sigma_{k+1}. \tag{2.14}$$

Proof: Let $X \in R^{m\times n}$ with $rank(X) = k$. Let $\tau_1, \cdots, \tau_n$ with $\tau_1 \geq \cdots \geq \tau_n \geq 0$ be the singular values of X. Since $A = X + (A - X)$ and $\tau_{k+1} = 0$, then $\sigma_{k+1} = |\tau_{k+1} - \sigma_{k+1}| \leq \|A - X\|_2$. For the matrix $A_k = U\tilde{\Sigma}V^T$ ($\tilde{\Sigma} = diag(\sigma_1, \cdots, \sigma_k, 0, \cdots, 0)$) we have

$$\|A - A_k\|_2 = \|U(\Sigma - \tilde{\Sigma})V^T\|_2 = \|\Sigma - \tilde{\Sigma}\|_2 = \sigma_{k+1}.$$

■

LS-problem: $\|Ax - b\|_2$=min! $\Rightarrow x_{LS} = A^+b$.

Perturbated LS-problem: $\|(A + E)y - (b + f)\|_2 = \min!$

$$\Rightarrow \quad y = (A + E)^+(b + f).$$

Lemma 7.2.13 *Let $A, E \in R^{m\times n}$ and $rank(A) = r$.*
(1) If $rank(A+E) > r$ then holds $\|(A+E)^+\|_2 \geq \frac{1}{\|E\|_2}$.
(2) If $rank(A+E) \leq r$ and $\|A^+\|_2\|E\|_2 < 1$ then $rank(A+E) = r$ and

$$\|(A+E)^+\|_2 \leq \frac{\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}.$$

Proof: Let $\tau_1 \geq \cdots \geq \tau_n$ be the singular values of $A+E$.

To (1): If τ_k is the smallest nonzero singular value, then $k \geq r+1$ because of $rank(A+E) > r$. By corollary 7.2.6 we have $\|E\|_2 = \|(A+E) - A\|_2 \geq \tau_{r+1} \geq \tau_k$ and therefore $\|(A+E)^+\|_2 = 1/\tau_k \geq 1/\|E\|_2$.

To (2): Let $\sigma_1 \geq \cdots \geq \sigma_n$ be the singular values of A, then $\sigma_r \neq 0$ because of $rank(A) = r$ and $\|A^+\|_2 = 1/\sigma_r$. Since $\|A^+\|_2\|E\|_2 < 1$ so $\|E\|_2 < \sigma_r$, and then by corollary 7.2.6 it must be $rank(A+E) \geq r$, so we have $rank(A+E) = r$. By Weyl's theorem (theorem 6.1.5) we have $\tau_r \geq \sigma_r - \|E\|_2$ and furthermore here $\sigma_r - \|E\|_2 > 0$, so one obtains

$$\|(A+E)^+\|_2 = 1/\tau_r \leq 1/(\sigma_r - \|E\|_2) = \|A^+\|_2/(1 - \|A^+\|_2\|E\|_2).$$

■

Lemma 7.2.14 *Let $A, E \in R^{m\times n}$, $b, f \in R^m$ and $x = A^+b$, $y = (A+E)^+(b+f)$ and $r = b - Ax$, then holds*

$$\begin{aligned} y - x &= [-(A+E)^+EA^+ + (A+E)^+(I - AA^+) \\ &\quad +(I - (A+E)^+(A+E)A^+]b + (A+E)^+f \\ &= -(A+E)^+Ex + (A+E)^+(A+E)^{+T}E^Tr \\ &\quad +(I - (A+E)^+(A+E))E^TA^{+T}x + (A+E)^+f. \end{aligned}$$

Proof: $y - x = [(A+E)^+ - A^+]b + (A+E)^+f$ and for $(A+E)^+ - A^+$ one has the decomposition

$$\begin{aligned} (A+E)^+ - A^+ &= -(A+E)^+EA^+ + (A+E)^+ - A^+ \\ &\quad +(A+E)^+(A+E-A)A^+ \\ &= -(A+E)^+EA^+ + (A+E)^+(I - AA^+) \\ &\quad -(I - (A+E)^+(A+E))A^+. \end{aligned}$$

Let $C := A+E$ and apply the generalized inverse to C we obtain $C^+ = C^+CC^+ = C^+C^{+^T}C^+$ and

$$A^T(I - AA^+) = A^T - A^TAA^+ = A^T - A^TA^{+^T}A^T = A^T - A^TA^{T^+}A^T = 0,$$

also $A^+ = A^TA^{+^T}A^+$ and $(I - C^+C)C^T = 0$. Hence it holds

$$C^+(I - AA^+) = C^+C^{+^T}E^T(I - AA^+)$$

and

$$(I - C^+C)A^+ = (I - C^+C)E^TA^{+^T}A^+.$$

If we substitute this into the second and third terms in the decomposition of $(A+E)^+ - A^+$ then we have the result ($r = (I - AA^+)b$, $x = A^+b$):

$$\begin{aligned} y - x &= [-(A+E)^+EA^+ + (A+E)^+(A+E)^{+^T}E^T(I - AA^T) \\ &\quad +(I - (A+E)^+(A+E))E^TA^{+^T}A^+]b + (A+E)^+f \\ &= -(A+E)^+Ex + (A+E)^+(A+E)^{+^T}E^Tr \\ &\quad +(I - (A+E)^+(A+E))E^TA^{+^T}x + (A+E)^+f \end{aligned}$$

Theorem 7.2.15 *Let $A, E \in R^{m\times n}$, $b, f \in R^m$, and $x = A^+b \neq 0$, $y = (A+E)^+(b+f)$ and $r = b - Ax$. If $rank(A) = r$, $rank(A+E) \leq r$ and $\|A^+\|_2\|E\|_2 < 1$, then holds*

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \frac{\|A\|_2\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}[2\frac{\|E\|_2}{\|A\|_2} + \frac{\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}\frac{\|E\|_2}{\|A\|_2}\frac{\|r\|_2}{\|x\|_2} + \frac{\|f\|_2}{\|A\|_2\|x\|_2}].$$

Proof: From Lemma 7.2.14 follows

$$\begin{aligned} \|y-x\|_2 &\leq \|(A+E)^+\|_2[\|E\|_2\|x\|_2 + \|(A+E)^+\|_2\|E\|_2\|r\|_2 + \|f\|_2] \\ &\quad +\|I - (A+E)^+(A+E)\|_2\|E\|_2\|A^+\|_2\|x\|_2. \end{aligned}$$

Since $I - (A+E)^+(A+E)$ is symmetric and it holds

$$(I - (A+E)^+(A+E))^2 = I - (A+E)^+(A+E).$$

From this follows $\|I - (A+E)^+(A+E)\|_2 = 1$, if $(A+E)^+(A+E) \neq I$. Together with the estimation of Lemma 7.2.13(2) we obtain

$$\|y-x\|_2 \leq \frac{\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}[2\|E\|_2\|x\|_2 + \|f\|_2 + \frac{\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}\|E\|_2\|r\|_2]$$

and

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \frac{\|A\|_2\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}[2\frac{\|E\|_2}{\|A\|_2} + \frac{\|f\|_2}{\|A\|_2\|x\|_2} + \frac{\|A^+\|_2}{1 - \|A^+\|_2\|E\|_2}\frac{\|E\|_2}{\|A\|_2}\frac{\|r\|_2}{\|x\|_2}].$$

■

7.3 Unsymmetric Lanczos Method

Suppose $A \in R^{n\times n}$ and that a nonsingular matrix X exists such that

$$X^{-1}AX = T = \begin{bmatrix} \alpha_1 & \gamma_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

Let

$$X = [x_1, \cdots, x_n] \text{ and } X^{-T} = Y = [y_1, \cdots, y_n].$$

Compare columns in $AX = XT$ and $A^TY = YT^T$ we find that

$$Ax_j = \gamma_{j-1}x_{j-1} + \alpha_j x_j + \beta_j x_{j+1}, \quad \gamma_0 x_0 \equiv 0$$

and

$$A^T y_j = \beta_{j-1}y_{j-1} + \alpha_j y_j + \gamma_j y_{j+1}, \quad \beta_0 y_0 \equiv 0$$

for $j = 1, \cdots, n-1$. These equations together with $Y^TX = I_n$ imply $\alpha_j = y_j^T A x_j$ and

$$\begin{aligned} \beta_j x_{j+1} &= \gamma_j \equiv (A - \alpha_j)x_j - \gamma_{j-1}x_{j-1} \\ \gamma_j y_{j+1} &= p_j \equiv (A - \alpha_j)^T y_j - \beta_{j-1}y_{j-1}. \end{aligned} \tag{3.1}$$

These is some flexibility in choosing the scale factors β_j and γ_j. A "canonical" choice is to set $\beta_j = \|\gamma_j\|_2$ and $\gamma_j = x_{j+1}^T p_j$ giving:

Biorthogonalization method of Lanczos:

$$\begin{array}{l} \text{Given } x_1, y_1 \in R^n \text{ with } x_1^T x_1 = y_1^T y_1 = 1. \\ \text{For } j = 1, \cdots, n-1, \\ \quad \alpha_j = y_j^T A x_j, \\ \quad r_j = (A - \alpha_j)x_j - \gamma_{j-1}x_{j-1} \quad (\gamma_0 x_0 \equiv 0), \\ \quad \beta_j = \|r_j\|_2. \\ \quad \text{If } \beta_j > 0 \text{ then } x_{j+1} = r_j/\beta_j \\ \quad p_j = (A - \alpha_j)^T y_j - \beta_{j-1}y_{j-1} \quad (\beta_0 y_0 \equiv 0), \\ \quad \gamma_j = x_{j+1}^T p_j, \text{ else stop;} \\ \quad \text{If } \gamma_j \neq 0 \text{ then } y_{j+1} = p_j/\gamma_j \text{ else stop;} \\ \quad \alpha_n = x_n^T A y_n \end{array} \tag{3.2}$$

Define $X_j = [x_1, \cdots, x_j]$, $Y_j = [y_1, \cdots, y_j]$ and T_j to be the leading $j \times j$ principal submatrix of T, it is easy to verify that

$$\begin{aligned} AX_j &= X_j T_j + \gamma_j e_j^T \\ A^T Y_j &= Y_j T_j^T + p_j e_j^T. \end{aligned} \tag{3.3}$$

Remark: (1) $p_j^T \gamma_j = \beta_j \gamma_j x_{j+1}^T y_{j+1} = \beta_j \gamma_j$ from (3.1).
(2) Break of the algorithm (3.2) occurs if $p_j^T \gamma_j = 0$:
(a) $\gamma_j = 0 \Rightarrow \beta_j = 0$. Then X_j is an invariant subspace of A (by (3.3)).
(b) $p_j = 0 \Rightarrow \gamma_j = 0$. Then Y_j is an invariant subspace of A^T (by (3.3)).

(c) $p_j^T \gamma_j = 0$ but $\|p_j\|\|\gamma_j\| \neq 0$, then (3.2) breaks down. We begin the algorithm (3.2) with a new starting vector.
(3) If $p_j^T \gamma_j$ is very small, then γ_j or β_j small. Hence y_{j+1} or x_{j+1} are large, so the algorithm (3.2) is unstable.

Definition 7.3.1 *An upper Hessenberg matrix $H = (h_{ij})$ is called unreducible, if $h_{i+1,i} \neq 0$, for $i = 1, \cdots, n-1$ (that is subdiagonal entries are nonzero). A tridiagonal matrix $T = (t_{ij})$ is called unreducible, if $t_{i,i-1} \neq 0$ for $i = 2, \cdots, n$ and $t_{i,i+1} \neq 0$ for $i = 1, \cdots, n-1$*

Theorem 7.3.2 *Let $A \in R^{n\times n}$. Then*
(1) If $x \neq 0$ so that $K[x_1, A, n] = [x_1, Ax_1, \cdots, A^{n-1}x_1]$ nonsingular and if X is a nonsingular matrix such that $K[x_1, A, n] = XR$, where R is an upper triangular matrix, then $H = X^{-1}AX$ is an upper unreducible Hessenberg matrix.

(2) Let X be a nonsingular matrix with first column x_1 and if $H = X^{-1}AX$ is an upper Hessenberg matrix, then holds

$$K[x_1, A, n] = XK[e_1, H, n] \equiv XR,$$

where R is an upper triangular matrix. Furthermore, if H is unreducible, then R is nonsingular.

(3) If $H = X^{-1}AX$ and $\tilde{H} = Y^{-1}AY$ where H and $\tilde{H}$ are both upper Hessenberg matrices, H is unreducible and the first columns x_1 and y_1 of X and Y, respectively, are linearly dependent, then $J = X^{-1}Y$ is an upper triangular matrix and $\tilde{H} = J^{-1}HJ$.

Proof: ad(1): Since $x_1, Ax_1, \cdots, A^{n-1}x_1$ are linearly independent, so $A^n x_1$ is the linear combination of $\{x_1, Ax_1, \cdots, A^{n-1}x_1\}$ i.e. there exists $c_0, \cdots, c_{n-1}$ such that

$$A^n x_1 = \sum_{i=0}^{n-1} c_i A^i x_1.$$

Let

$$C = \begin{bmatrix} 0 & \cdots & 0 & c_0 \\ 1 & \ddots & & c_1 \\ & \ddots & 0 & \vdots \\ 0 & & 1 & c_{n-1} \end{bmatrix}.$$

Then we have $K[x_1, A, n]C = [Ax_1, A^2x_1, \cdots, A^{n-1}x_1, A^n x_1] = AK[x_1, A, n]$. Thus $XRC = AXR$. We then have

$$X^{-1}AX = RCR^{-1} = H \text{ unreducible Hessenberg matrix.}$$

ad(2): From $A = XHX^{-1}$ follows that $A^i x_1 = XH^i X^{-1}x_1 = XH^i e_1$. Then

$$\begin{aligned} K[x_1, A, n] &= [x_1, Ax_1, \cdots, A^{n-1}x_1] = [Xe_1, XHe_1, \cdots, XH^{n-1}e_1] \\ &= X[e_1, He_1, \cdots, H^{n-1}e_1]. \end{aligned}$$

If H is upper Hessenberg, then $R = [e_1, He_1, \cdots, H^{n-1}e_1]$ is upper triangular. If H is unreducible upper Hessenberg, then R is nonsingular, since $r_{11} = 1$, $r_{22} = h_{21}$, $r_{33} = h_{21}h_{32}, \cdots$, and so on.

ad(3): Let $y_1 = \lambda x_1$. We apply (2) to the matrix H. It follows $K[x_1, A, n] = XR_1$. Applying (2) to $\tilde{H}$, we also have $K[y_1, A, n] = YR_2$. Here R_1 and R_2 are upper triangular. Since $y_1 = \lambda x_1$, so

$$\lambda K[x_1, A, n] = \lambda XR_1 = YR_2.$$

Since R_1 is nonsingular, by (2) we have R_2 is nonsingular and $X^{-1}Y = \lambda R_1 R_2^{-1} = J$ (upper triangular). So

$$\tilde{H} = Y^{-1}AY = (Y^{-1}X)X^{-1}AX(X^{-1}Y) = J^{-1}HJ.$$

■

Theorem 7.3.3 *Let $A \in R^{n\times n}$, $x, y \in R^n$ with $K[x, A, n]$ and $K[y, A^T, n]$ nonsingular. Then*

(1) If $B = K[y, A^T, n]^T K[x, A, n] = (y^T A^{i+j-2}x)_{i,j=1,\cdots,n}$ has a decomposition $B = LDL^T$, where L is a lower triangular with $l_{ii} = 1$ and D is diagonal (that is all principal determinants of B are nonzero) and if $X = K[x, A, n]L^{-1}$, then $T = X^{-1}AX$ is an unreducible tridiagonal matrix.

(2) Let X, Y be nonsingualr with
(a) $T = X^{-1}AX$, $\tilde{T} = Y^{-1}AY$ unreducible tridiagonal,
(b) the first column of X and Y are linearly dependent,
(c) the first row of X and Y are linearly dependent.
Then $X^{-1}Y = D$ diagonal and $\tilde{T} = D^{-1}TD$.

(3) If $T = X^{-1}AX$ is unreducible tridiagonal, x is the first column of X and Y is the first row of X^{-1}, then

$$B = K[y, A^T, n]^T K[x, A, n]$$

has a LDL^T decomposition.

Proof: ad(1):

$$X = K[x, A, n]L^{-T} \Rightarrow XL^T = K[x, A, n]. \tag{3.4}$$

So the first column of X is x. From $B = LDL^T$ follows

$$K[y, A^T, n]^T = LDL^T K[x, A, n]^{-1}$$

and then

$$K[y, A^T, n] = K[x, A, n]^{-T} LDL^T = X^{-T}DL^T. \tag{3.5}$$

Apply theorem 7.3.2(1) to (3.4):

$$X^{-1}AX \text{ unreducible upper Hessenberg.}$$

Apply theorem 7.3.2(1) to (3.5):

$$X^T A^T X^{-T} = (X^{-1}AX)^T \text{ unreducible upper Hessenberg.}$$

So $X^{-1}AX$ is an unreducible tridiagonal matrix.

ad(2): T and $\tilde{T}$ are unreducible upper Hessenberg, by theorem 7.3.2(3) we have $X^{-1}Y$ upper triangular on the other hand. Since

$$\left.\begin{array}{rcll} T^T & = & X^T A^T X^{-T} & \text{upper Hessenberg} \\ \tilde{T}^T & = & Y^T A^T Y^{-T} & \text{upper Hessenberg} \end{array}\right\} \text{unreducible,}$$

then by theorem 7.3.2(3) we also have

$$Y^T X^{-T} = (X^{-1}Y)^T \text{ upper triangular.}$$

Thus $X^{-1}Y$ is upper triangular, also lower triangular so the matrix $X^{-1}Y$ is diagonal.
ad(3): exercise! ■

Chapter 8

Arnoldi Method

8.1 Arnoldi decompositions

Suppose that the columns of K_{k+1} are linearly independent and let

$$K_{k+1} = U_{k+1} R_{k+1}$$

be the QR factorization of K_{k+1}.

Theorem 8.1.1 *Let $\|u_1\|_2 = 1$ and the columns of $K_{k+1}(A, u_1)$ be linearly independent. Let $U_{k+1} = [\ u_1\ \ \cdots\ \ u_{k+1}\ ]$ be the Q-factor of K_{k+1}. Then there is a $(k+1) \times k$ unreduced upper Hessenberg matrix*

$$\hat{H}_k \equiv \left[\begin{array}{cccc} \hat{h}_{11} & \cdots & \cdots & \hat{h}_{1k} \\ \hat{h}_{21} & \hat{h}_{22} & \cdots & \hat{h}_{2k} \\ & \ddots & \ddots & \vdots \\ & & \hat{h}_{k,k-1} & \hat{h}_{kk} \\ \hline & & & \hat{h}_{k+1,k} \end{array}\right] \qquad \textit{with} \qquad \hat{h}_{i+1,i} \neq 0$$

such that

$$AU_k = U_{k+1} \hat{H}_k. \tag{8.1.1}$$

Conversely, if U_{k+1} is orthonormal and satisfies (8.1.1), where $\hat{H}_k$ is a $(k+1)\times k$ unreduced upper Hessenberg matrix, then U_{k+1} is the Q-factor of $K_{k+1}(A, u_1)$.

Proof: ("$\Rightarrow$") Let $K_k = U_k R_k$ be the QR factorization and $S_k = R_k^{-1}$. Then

$$AU_k = AK_k S_k = K_{k+1} \begin{bmatrix} 0 \\ S_k \end{bmatrix} = U_{k+1} R_{k+1} \begin{bmatrix} 0 \\ S_k \end{bmatrix} = U_{k+1} \hat{H}_k,$$

where

$$\hat{H}_k = R_{k+1} \begin{bmatrix} 0 \\ S_k \end{bmatrix}.$$

It implies that $\hat{H}_k$ is a $(k+1) \times k$ Hessenberg matrix and

$$\hat{h}_{i+1,i} = r_{i+1,i+1} s_{ii} = \frac{r_{i+1,i+1}}{r_{ii}}.$$

Thus by the nonsingularity of R_k, $\hat{H}_k$ is unreduced.

("$\Leftarrow$") If $k = 1$, then

$$Au_1 = \hat{h}_{11} u_1 + \hat{h}_{21} u_2.$$

It follows that

$$K_2(A, u_1) = [u_1 \ Au_1] = [u_1 \ u_2] \begin{bmatrix} 1 & \hat{h}_{11} \\ 0 & \hat{h}_{21} \end{bmatrix}.$$

Since $[u_1 \ u_2]$ is orthonormal, $[u_1 \ u_2]$ is the Q-factor of K_2.

Assume U_k is the Q-factor of $K_k(A, u_1)$, i.e.

$$K_k(A, u_1) = U_k R_k,$$

where R_k is upper triangular. If we partition

$$\hat{H}_k = \begin{bmatrix} \hat{H}_{k-1} & \hat{h}_k \\ 0 & \hat{h}_{k+1,k} \end{bmatrix},$$

then from (8.1.1)

$$\begin{aligned} K_{k+1}(A, u_1) &= \begin{bmatrix} K_k(A, u_1) & Au_k \end{bmatrix} \\ &= \begin{bmatrix} U_k R_k & U_k \hat{h}_k + \hat{h}_{k+1,k} u_{k+1} \end{bmatrix} \\ &= \begin{bmatrix} U_k & u_{k+1} \end{bmatrix} \begin{bmatrix} R_k & \hat{h}_k \\ 0 & \hat{h}_{k+1,k} \end{bmatrix}. \end{aligned}$$

Hence U_{k+1} is the Q-factor of K_{k+1}. ■

Definition 8.1.1 *Let $U_{k+1} \in \mathbb{C}^{n \times (k+1)}$ be orthonormal. If there is a $(k+1) \times k$ unreduced upper Hessenberg matrix $\hat{H}_k$ such that*

$$AU_k = U_{k+1} \hat{H}_k, \tag{8.1.2}$$

then (8.1.2) is called an Arnoldi decomposition of order k. If $\hat{H}_k$ is reduced, we say the Arnoldi decomposition is reduced.

Partition

$$\hat{H}_k = \begin{bmatrix} H_k \\ \hat{h}_{k+1,k} e_k^T \end{bmatrix},$$

and set

$$\beta_k = \hat{h}_{k+1,k}.$$

Then (8.1.2) is equivalent to

$$AU_k = U_k H_k + \beta_k u_{k+1} e_k^T.$$

Theorem 8.1.2 *Suppose the Krylov sequence $K_{k+1}(A, u_1)$ does not terminate at $k+1$. Then up to scaling of the columns of U_{k+1}, the Arnoldi decomposition of K_{k+1} is unique.*

Proof: Since the Krylov sequence $K_{k+1}(A, u_1)$ does not terminate at $k+1$, the columns of $K_{k+1}(A, u_1)$ are linearly independent. By Theorem 8.1.1, there is an unreduced matrix H_k and $\beta_k \neq 0$ such that

$$AU_k = U_k H_k + \beta_k u_{k+1} e_k^T, \tag{8.1.3}$$

where $U_{k+1} = [U_k \; u_{k+1}]$ is an orthonormal basis for $\mathcal{K}_{k+1}(A, u_1)$. Suppose there is another orthonormal basis $\tilde{U}_{k+1} = [\tilde{U}_k \; \tilde{u}_{k+1}]$ for $\mathcal{K}_{k+1}(A, u_1)$, unreduced matrix $\tilde{H}_k$ and $\tilde{\beta}_k \neq 0$ such that

$$A\tilde{U}_k = \tilde{U}_k \tilde{H}_k + \tilde{\beta}_k \tilde{u}_{k+1} e_k^T.$$

Then we claim that

$$\tilde{U}_k^H u_{k+1} = 0.$$

For otherwise there is a column $\tilde{u}_j$ of $\tilde{U}_k$ such that

$$\tilde{u}_j = \alpha u_{k+1} + U_k a, \quad \alpha \neq 0.$$

Hence

$$A\tilde{u}_j = \alpha A u_{k+1} + AU_k a$$

which implies that $A\tilde{u}_j$ contains a component along $A^{k+1}u_1$. Since the Krylov sequence $K_{k+1}(A, u_1)$ does not terminate at $k+1$, we have

$$\mathcal{K}_{k+2}(A, u_1) \neq \mathcal{K}_{k+1}(A, u_1).$$

Therefore, $A\tilde{u}_j$ lies in $\mathcal{K}_{k+2}(A, u_1)$ but not in $\mathcal{K}_{k+1}(A, u_1)$ which is a contradiction.

Since U_{k+1} and $\tilde{U}_{k+1}$ are orthonormal bases for $\mathcal{K}_{k+1}(A, u_1)$ and $\tilde{U}_k^H u_{k+1} = 0$, it follows that

$$\mathcal{R}(U_k) = \mathcal{R}(\tilde{U}_k) \quad \text{and} \quad U_k^H \tilde{u}_{k+1} = 0,$$

that is

$$U_k = \tilde{U}_k Q$$

for some unitary matrix Q. Hence

$$A(\tilde{U}_k Q) = (\tilde{U}_k Q)(Q^H \tilde{H}_k Q) + \tilde{\beta}_k \tilde{u}_{k+1}(e_k^T Q),$$

or

$$AU_k = U_k(Q^H \tilde{H}_k Q) + \tilde{\beta}_k \tilde{u}_{k+1} e_k^T Q. \tag{8.1.4}$$

On premultiplying (8.1.3) and (8.1.4) by U_k^H, we obtain

$$H_k = U_k^H A U_k = Q^H \tilde{H}_k Q.$$

Similarly, premultiplying by u_{k+1}^H, we obtain

$$\beta_k e_k^T = u_{k+1}^H A U_k = \tilde{\beta}_k (u_{k+1}^H \tilde{u}_{k+1}) e_k^T Q.$$

It follows that the last row of Q is $\omega_k e_k^T$, where $|\omega_k| = 1$. Since the norm of the last column of Q is one, the last column of Q is $\omega_k e_k$. Since H_k is unreduced, it follows from the implicit Q theorem that

$$Q = \mathrm{diag}(\omega_1, \cdots, \omega_k), \quad |\omega_j| = 1,\ j = 1, \ldots, k.$$

Thus up to column scaling $U_k = \tilde{U}_k Q$ is the same as $\tilde{U}_k$. Subtracting (8.1.4) from (8.1.3), we find that

$$\beta_k u_{k+1} = \omega_k \tilde{\beta}_k \tilde{u}_{k+1}$$

so that up to scaling u_{k+1} and $\tilde{u}_{k+1}$ are the same. ■

Theorem 8.1.3 *Let the orthonormal matrix U_{k+1} satisfy*

$$A U_k = U_{k+1} \hat{H}_k,$$

where $\hat{H}_k$ is Hessenberg. Then $\hat{H}_k$ is reduced if and only if $\mathcal{R}(U_k)$ contains an eigenspace of A.

Proof: ("⇒") Suppose that $\hat{H}_k$ is reduced, say that $h_{j+1,j} = 0$. Partition

$$\hat{H}_k = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix} \quad \text{and} \quad U_k = \begin{bmatrix} U_{11} & U_{12} \end{bmatrix},$$

where H_{11} is an $j \times j$ matrix and U_{11} is consisted the first j columns of U_{k+1}. Then

$$A \begin{bmatrix} U_{11} & U_{12} \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} & u_{k+1} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}.$$

It implies that

$$A U_{11} = U_{11} H_{11}$$

so that U_{11} is an eigenbasis of A.

("⇐") Suppose that A has an eigenspace that is a subset of $\mathcal{R}(U_k)$ and $\hat{H}_k$ is unreduced. Let $(\lambda, U_k w)$ for some w be an eigenpair of A. Then

$$\begin{aligned} 0 &= (A - \lambda I) U_k w = (U_{k+1} \hat{H}_k - \lambda U_k) w \\ &= \left(U_{k+1} \hat{H}_k - \lambda U_{k+1} \begin{bmatrix} I \\ 0 \end{bmatrix} \right) w = U_{k+1} \hat{H}_\lambda w, \end{aligned}$$

where

$$\hat{H}_\lambda = \begin{bmatrix} H_k - \lambda I \\ h_{k+1,k} e_k^T \end{bmatrix}.$$

Since $\hat{H}_\lambda$ is unreduced, the matrix $U_{k+1}\hat{H}_\lambda$ is of full column rank. It follows that $w = 0$ which is a contradiction. ∎

Write the k-th column of the Arnoldi decomposition

$$AU_k = U_k H_k + \beta_k u_{k+1} e_k^T,$$

in the form

$$Au_k = U_k h_k + \beta_k u_{k+1}.$$

Then from the orthonormality of U_{k+1}, we have

$$h_k = U_k^H A u_k.$$

Since

$$\beta_k u_{k+1} = Au_k - U_k h_k$$

and $\|u_{k+1}\|_2 = 1$, we must have

$$\beta_k = \|Au_k - U_k h_k\|_2$$

and

$$u_{k+1} = \beta_k^{-1}(Au_k - U_k h_k).$$

Algorithm 8.1.1 (Arnoldi process)

1. *For* $k = 1, 2, \ldots$
2. $\quad h_k = U_k^H A u_k.$
3. $\quad v = Au_k - U_k h_k$
4. $\quad \beta_k = h_{k+1,k} = \|v\|_2$
5. $\quad u_{k+1} = v/\beta_k$
6. $\quad \hat{H}_k = \begin{bmatrix} \hat{H}_{k-1} & h_k \\ 0 & h_{k+1,k} \end{bmatrix}$
7. *end for* k

The computation of u_{k+1} is actually a form of the well-known Gram-Schmidt algorithm. In the presence of inexact arithmetic cancelation in statement 3 can cause it to fail to produce orthogonal vectors. The cure is process called reorthogonalization.

Algorithm 8.1.2 (Reorthogonalized Arnoldi process)

For $k = 1, 2, \ldots$
$\quad h_k = U_k^H A u_k.$
$\quad v = Au_k - U_k h_k.$
$\quad w = U_k^H v.$
$\quad h_k = h_k + w.$
$\quad v = v - U_k w.$
$\quad \beta_k = h_{k+1,k} = \|v\|_2$
$\quad u_{k+1} = v/\beta_k$
$\quad \hat{H}_k = \begin{bmatrix} \hat{H}_{k-1} & h_k \\ 0 & h_{k+1,k} \end{bmatrix}$
end for k

Let $y_i^{(k)}$ be an eigenvector of H_k associated with the eigenvalue $\lambda_i^{(k)}$ and $x_i^{(k)} = U_k y_i^{(k)}$ the Ritz approximate eigenvector.

Theorem 8.1.4

$$(A - \lambda_i^{(k)} I)x_i^{(k)} = h_{k+1,k} e_k^T y_i^{(k)} u_{k+1}.$$

and therefore,

$$\|(A - \lambda_i^{(k)} I)x_i^{(k)}\|_2 = |h_{k+1,k}|\, |e_k^T y_i^{(k)}|.$$

8.2 Krylov decompositions

Definition 8.2.1 *Let $u_1, u_2, \ldots, u_{k+1}$ be linearly independent and let $U_k = [u_1 \ \cdots \ u_k]$.*

$$AU_k = U_k B_k + u_{k+1} b_{k+1}^H$$

is called a Krylov decomposition of order k. $\mathcal{R}(U_{k+1})$ is called the space spanned by the decomposition. Two Krylov decompositions spanning the same spaces are said to be equivalent.

Let $[V \ v]^H$ be any left inverse for U_{k+1}. Then it follows that

$$B_k = V^H A U_k \quad \text{and} \quad b_{k+1}^H = v^H A U_k.$$

In particular, B_k is a Rayleigh quotient of A.

Let

$$AU_k = U_k B_k + u_{k+1} b_{k+1}^H$$

be a Krylov decomposition and Q be nonsingular. That is

$$AU_k = U_{k+1} \hat{B}_k \quad \text{with} \quad \hat{B}_k = \begin{bmatrix} B_k \\ b_{k+1}^H \end{bmatrix}. \tag{8.2.5}$$

Then we get an equivalent Krylov decomposition of (8.2.5) in the form

$$\begin{aligned} A(U_k Q) &= \left(U_{k+1} \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \right) \left(\begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix}^{-1} \hat{B}_k Q \right) \\ &= \begin{bmatrix} U_k Q & u_{k+1} \end{bmatrix} \begin{bmatrix} Q^{-1} B_k Q \\ b_{k+1}^H Q \end{bmatrix} \\ &= (U_k Q)(Q^{-1} B Q) + u_{k+1}(b_{k+1}^H Q). \end{aligned} \tag{8.2.6}$$

The two Krylov decompositions (8.2.5) and (8.2.6) are said to be similar.

Let

$$\gamma \tilde{u}_{k+1} = u_{k+1} - U_k a.$$

Since $u_1, \ldots, u_k, u_{k+1}$ are linearly independent, we have $\gamma \neq 0$. Then it follows that

$$AU_k = U_k(B_k + a b_{k+1}^H) + \tilde{u}_{k+1}(\gamma b_{k+1}^H).$$

Since $\mathcal{R}([U_k \ u_{k+1}]) = \mathcal{R}([U_k \ \tilde{u}_{k+1}])$, this Krylov decomposition is equivalent to (8.2.5).

Theorem 8.2.1 *Every Krylov decomposition is equivalent to a (possibly reduced) Arnoldi decomposition.*

Proof: Let

$$AU = UB + ub^H$$

be a Krylov decomposition and let

$$U = \tilde{U}R$$

be the QR factorization of U. Then

$$A\tilde{U} = A(UR^{-1}) = (UR^{-1})(RBR^{-1}) + u(b^H R^{-1}) \equiv \tilde{U}\tilde{B} + u\tilde{b}^H$$

is an equivalent decomposition. Let

$$\tilde{u} = \gamma^{-1}(u - Ua)$$

be a vector with $\|\tilde{u}\|_2 = 1$ such that $U^H\tilde{u} = 0$. Then

$$A\tilde{U} = \tilde{U}(\tilde{B} + a\tilde{b}^H) + \tilde{u}(\gamma\tilde{b}^H) \equiv \tilde{U}\hat{B} + \tilde{u}\hat{b}^H$$

is an equivalent orthonormal Krylov decomposition. Let Q be a unitary matrix such that

$$\hat{b}^H Q = \|\hat{b}\|_2 e_k^T$$

and $Q^H\hat{B}Q$ is upper Hessenberg. Then the equivalent decomposition

$$A\hat{U} \equiv A(\tilde{U}Q) = (\tilde{U}Q)(Q^H\hat{B}Q) + \tilde{u}(\hat{b}^H Q) \equiv \hat{U}\bar{B} + \|\hat{b}\|_2\hat{u}e_k^T$$

is a possibly reduced Arnoldi decomposition where

$$\hat{U}^H\hat{u} = Q^H\tilde{U}^H\tilde{u} = Q^H R^{-H}U^H\tilde{u} = 0.$$

■

8.2.1 Reduction to Arnoldi form

Let

$$AU = UB + ub^H$$

be the Krylov decomposition with $B \in \mathbb{C}^{k\times k}$. Let H_1 be a Householder transformation such that

$$b^H H_1 = \beta e_k.$$

Reduce $H_1^H B H_1$ to Hessenberg form as the following illustration:

$$B := \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix} \Rightarrow B := BH_2 = \begin{bmatrix} \otimes & \otimes & \otimes & \times \\ \otimes & \otimes & \otimes & \times \\ \otimes & \otimes & \otimes & \times \\ 0 & 0 & \otimes & \times \end{bmatrix}$$

$$\Rightarrow \quad B := H_2^H B = \begin{bmatrix} + & + & + & + \\ + & + & + & + \\ + & + & + & + \\ 0 & 0 & \otimes & \times \end{bmatrix} \Rightarrow B := BH_3 = \begin{bmatrix} \oplus & \oplus & + & + \\ \oplus & \oplus & + & + \\ 0 & \oplus & + & + \\ 0 & 0 & \otimes & \times \end{bmatrix}$$

$$\Rightarrow \quad B := H_3^H B = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & \oplus & + & + \\ 0 & 0 & \otimes & \times \end{bmatrix}$$

Let

$$Q = H_1 H_2 \cdots H_{k-1}.$$

Then $Q^H B Q$ is upper Hessenberg and

$$b^H Q = (b^H H_1)(H_2 \cdots H_{k-1}) = \beta e_k^T (H_2 \cdots H_{k-1}) = \beta e_k^T.$$

Therefore, the Krylov decomposition

$$A(UQ) = (UQ)(Q^H B Q) + \beta u e_k^T \tag{8.2.7}$$

is an Arnoldi decomposition.

8.3 The implicitly restarted Arnoldi method

Let

$$AU_k = U_k H_k + \beta_k u_{k+1} e_k^T$$

be an Arnoldi decomposition.

- In principle, we can keep expanding the Arnoldi decomposition until the Ritz pairs have converged.
- Unfortunately, it is limited by the amount of memory to storage of U_k.
- Restarted the Arnoldi process once k becomes so large that we cannot store U_k.
  - Implicitly restarting method
  - Krylov-Schur decomposition
- Choose a new starting vector for the underlying Krylov sequence
- A natural choice would be a linear combination of Ritz vectors that we are interested in.

8.3.1 Filter polynomials

Assume A has a complete system of eigenpairs (λ_i, x_i) and we are interested in the first k of these eigenpairs. Expand u_1 in the form

$$u_1 = \sum_{i=1}^{k} \gamma_i x_i + \sum_{i=k+1}^{n} \gamma_i x_i.$$

If p is any polynomial, we have

$$p(A)u_1 = \sum_{i=1}^{k} \gamma_i p(\lambda_i) x_i + \sum_{i=k+1}^{n} \gamma_i p(\lambda_i) x_i.$$

- Choose p so that the values $p(\lambda_i)$ $(i = k+1, \ldots, n)$ are small compared to the values $p(\lambda_i)$ $(i = 1, \ldots, k)$.
- Then $p(A)u_1$ is rich in the components of the x_i that we want and deficient in the ones that we do not want.
- p is called a filter polynomial.
- Suppose we have Ritz values $\mu_1, \ldots, \mu_m$ and $\mu_{k+1}, \ldots, \mu_m$ are not interesting. Then take

$$p(t) = (t - \mu_{k+1}) \cdots (t - \mu_m).$$

8.3.2 Implicitly restarted Arnoldi

Let

$$AU_m = U_m H_m + \beta_m u_{m+1} e_m^T \tag{8.3.8}$$

be an Arnoldi decomposition with order m. Choose a filter polynomial p of degree $m - k$ and use the implicit restarting process to reduce the decomposition to a decomposition

$$A\tilde{U}_k = \tilde{U}_k \tilde{H}_k + \tilde{\beta}_k \tilde{u}_{k+1} e_k^T$$

of order k with starting vector $p(A)u_1$.

Let $\kappa_1, \ldots, \kappa_m$ be eigenvalues of H_m and suppose that $\kappa_1, \ldots, \kappa_{m-k}$ correspond to the part of the spectrum we are not interested in. Then take

$$p(t) = (t - \kappa_1)(t - \kappa_2) \cdots (t - \kappa_{m-k}).$$

The starting vector $p(A)u_1$ is equal to

$$\begin{aligned} p(A)u_1 &= (A - \kappa_{m-k} I) \cdots (A - \kappa_2 I)(A - \kappa_1 I) u_1 \\ &= (A - \kappa_{m-k} I) \left[\cdots \left[(A - \kappa_2 I) \left[(A - \kappa_1 I) u_1 \right] \right] \right]. \end{aligned}$$

In the first, we construct an Arnoldi decomposition with starting vector $(A - \kappa_1 I)u_1$. From (8.3.8), we have

$$\begin{aligned}(A - \kappa_1 I)U_m &= U_m(H_m - \kappa_1 I) + \beta_m u_{m+1} e_m^T \\ &= U_m Q_1 R_1 + \beta_m u_{m+1} e_m^T,\end{aligned} \tag{8.3.9}$$

where

$$H_m - \kappa_1 I = Q_1 R_1$$

is the QR factorization of $H_m - \kappa_1 I$. Postmultiplying by Q_1, we get

$$(A - \kappa_1 I)(U_m Q_1) = (U_m Q_1)(R_1 Q_1) + \beta_m u_{m+1}(e_m^T Q_1).$$

It implies that

$$AU_m^{(1)} = U_m^{(1)} H_m^{(1)} + \beta_m u_{m+1} b_{m+1}^{(1)H},$$

where

$$U_m^{(1)} = U_m Q_1, \quad H_m^{(1)} = R_1 Q_1 + \kappa_1 I, \quad b_{m+1}^{(1)H} = e_m^T Q_1.$$

($H_m^{(1)}$: one step of single shifted QR algorithm)

Theorem 8.3.1 *Let H_m be an unreduced Hessenberg matrix. Then $H_m^{(1)}$ has the form*

$$H_m^{(1)} = \begin{bmatrix} \hat{H}_m^{(1)} & \hat{h}_{12} \\ 0 & \kappa_1 \end{bmatrix},$$

where $\hat{H}_m^{(1)}$ is unreduced.

Proof: Let

$$H_m - \kappa_1 I = Q_1 R_1$$

be the QR factorization of $H_m - \kappa_1 I$ with

$$Q_1 = G(1, 2, \theta_1) \cdots G(m-1, m, \theta_{m-1})$$

where $G(i, i+1, \theta_i)$ for $i = 1, \ldots, m-1$ are Given rotations. Since H_m is unreduced upper Hessenberg, i.e., the subdiagonal elements of H_m are nonzero, we get

$$\theta_i \neq 0 \quad \text{for} \quad i = 1, \ldots, m-1 \tag{8.3.10}$$

and

$$(R_1)_{ii} \neq 0 \quad \text{for} \quad i = 1, \ldots, m-1. \tag{8.3.11}$$

Since κ_1 is an eigenvalue of H_m, we have that $H_m - \kappa_1 I$ is singular and then

$$(R_1)_{mm} = 0. \tag{8.3.12}$$

Using the results of (8.3.10), (8.3.11) and (8.3.12), we get

$$\begin{aligned}H_m^{(1)} &= R_1 Q_1 + \kappa_1 I = R_1 G(1, 2, \theta_1) \cdots G(m-1, m, \theta_{m-1}) + \kappa_1 I \\ &= \begin{bmatrix} \hat{H}_m^{(1)} & \hat{h}_{12} \\ 0 & \kappa_1 \end{bmatrix},\end{aligned}$$

where $\hat{H}_m^{(1)}$ is unreduced. ■

Remark 8.3.1

- $U_m^{(1)}$ *is orthonormal.*
- *Since* H_m *is upper Hessenberg and* Q_1 *is the* Q*-factor of the* QR *factorization of* $H_m - \kappa_1 I$*, it implies that* Q_1 *and* $H_m^{(1)}$ *are also upper Hessenberg.*
- *The vector* $b_{m+1}^{(1)H} = e_m^T Q_1$ *has the form*

$$b_{m+1}^{(1)H} = \begin{bmatrix} 0 & \cdots & 0 & q_{m-1,m}^{(1)} & q_{m,m}^{(1)} \end{bmatrix};$$

  i.e., only the last two components of $b_{m+1}^{(1)}$ *are nonzero.*
- *For on postmultiplying (8.3.9) by* e_1*, we get*

$$(A - \kappa_1 I)u_1 = (A - \kappa_1 I)(U_m e_1) = U_m^{(1)} R_1 e_1 = r_{11}^{(1)} u_1^{(1)}.$$

  Since H_m *is unreduced,* $r_{11}^{(1)}$ *is nonzero. Therefore, the first column of* $U_m^{(1)}$ *is a multiple of* $(A - \kappa_1 I)u_1$*.*
- *By the definition of* $H_m^{(1)}$*, we get*

$$Q_1 H_m^{(1)} Q_1^H = Q_1(R_1 Q_1 + \kappa_1 I)Q_1^H = Q_1 R_1 + \kappa_1 I = H_m.$$

  Therefore, $\kappa_1, \kappa_2, \ldots, \kappa_m$ *are also eigenvalues of* $H_m^{(1)}$*.*

Similarly,

$$\begin{aligned} (A - \kappa_2 I)U_m^{(1)} &= U_m^{(1)}(H_m^{(1)} - \kappa_2 I) + \beta_m u_{m+1} b_{m+1}^{(1)H} \\ &= U_m^{(1)} Q_2 R_2 + \beta_m u_{m+1} b_{m+1}^{(1)H}, \end{aligned} \tag{8.3.13}$$

where

$$H_m^{(1)} - \kappa_2 I = Q_2 R_2$$

is the QR factorization of $H_m^{(1)} - \kappa_2 I$ with upper Hessenberg matrix Q_2. Postmultiplying by Q_2, we get

$$(A - \kappa_2 I)(U_m^{(1)} Q_2) = (U_m^{(1)} Q_2)(R_2 Q_2) + \beta_m u_{m+1}(b_{m+1}^{(1)H} Q_2).$$

It implies that

$$AU_m^{(2)} = U_m^{(2)} H_m^{(2)} + \beta_m u_{m+1} b_{m+1}^{(2)H},$$

where

$$U_m^{(2)} \equiv U_m^{(1)} Q_2$$

is orthonormal,

$$H_m^{(2)} \equiv R_2 Q_2 + \kappa_2 I = \left[\begin{array}{c|cc} H_{m-2}^{(2)} & * & * \\ \hline & \kappa_2 & * \\ & & \kappa_1 \end{array}\right]$$

is upper Hessenberg with unreduced matrix $H_{m-2}^{(2)}$ and

$$\begin{aligned} b_{m+1}^{(2)H} &\equiv b_{m+1}^{(1)H} Q_2 = q_{m-1,m}^{(1)} e_{m-1}^H Q_2 + q_{m,m}^{(1)} e_m^T Q_2 \\ &= \begin{bmatrix} 0 & \cdots & 0 & \times & \times & \times \end{bmatrix}. \end{aligned}$$

For on postmultiplying (8.3.13) by e_1, we get

$$(A - \kappa_2 I)u_1^{(1)} = (A - \kappa_2 I)(U_m^{(1)} e_1) = U_m^{(2)} R_2 e_1 = r_{11}^{(2)} u_1^{(2)}.$$

Since $H_m^{(1)}$ is unreduced, $r_{11}^{(2)}$ is nonzero. Therefore, the first column of $U_m^{(2)}$ is a multiple of $(A - \kappa_2 I)u_1^{(1)} = 1/r_{11}^{(1)}(A - \kappa_2 I)(A - \kappa_1 I)u_1$.

Repeating this process with $\kappa_3, \ldots, \kappa_{m-k}$, the result will be a Krylov decomposition

$$AU_m^{(m-k)} = U_m^{(m-k)} H_m^{(m-k)} + \beta_m u_{m+1} b_{m+1}^{(m-k)H}$$

with the following properties

i. $U_m^{(m-k)}$ is orthonormal.

ii. $H_m^{(m-k)}$ is upper Hessenberg.

iii. The first $k-1$ components of $b_{m+1}^{(m-k)H}$ are zero.

iv. The first column of $U_m^{(m-k)}$ is a multiple of $(A - \kappa_1 I) \cdots (A - \kappa_{m-k} I)u_1$.

Corollary 8.3.1 *Let $\kappa_1, \ldots, \kappa_m$ be eigenvalues of H_m. If the implicitly restarted QR step is performed with shifts $\kappa_1, \ldots, \kappa_{m-k}$, then the matrix $H_m^{(m-k)}$ has the form*

$$H_m^{(m-k)} = \begin{bmatrix} H_{kk}^{(m-k)} & H_{k,m-k}^{(m-k)} \\ 0 & T^{(m-k)} \end{bmatrix},$$

where $T^{(m-k)}$ is an upper triangular matrix with Ritz value $\kappa_1, \ldots, \kappa_{m-k}$ on its diagonal.

For $k = 3$ and $m = 6$,

$$\begin{aligned} & A \left[\begin{array}{ccc|ccc} u & u & u & u & u & u \end{array}\right] \\ = & \left[\begin{array}{ccc|ccc} u & u & u & u & u & u \end{array}\right] \left[\begin{array}{ccc|ccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ \hline 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{array}\right] \\ & + u \left[\begin{array}{ccc|ccc} 0 & 0 & q & q & q & q \end{array}\right]. \end{aligned}$$

Therefore, the first k columns of the decomposition can be written in the form

$$AU_k^{(m-k)} = U_k^{(m-k)} H_{kk}^{(m-k)} + h_{k+1,k} u_{k+1}^{(m-k)} e_k^T + \beta_k q_{mk} u_{m+1} e_k^T,$$

where $U_k^{(m-k)}$ consists of the first k columns of $U_m^{(m-k)}$, $H_{kk}^{(m-k)}$ is the leading principal submatrix of order k of $H_m^{(m-k)}$, and q_{km} is from the matrix $Q = Q_1 \cdots Q_{m-k}$. Hence if we set

$$\begin{aligned} \tilde{U}_k &= U_k^{(m-k)}, \\ \tilde{H}_k &= H_{kk}^{(m-k)}, \\ \tilde{\beta}_k &= \|h_{k+1,k} u_{k+1}^{(m-k)} + \beta_k q_{mk} u_{m+1}\|_2, \\ \tilde{u}_{k+1} &= \tilde{\beta}_k^{-1}(h_{k+1,k} u_{k+1}^{(m-k)} + \beta_k q_{mk} u_{m+1}), \end{aligned}$$

then

$$A\tilde{U}_k = \tilde{U}_k \tilde{H}_k + \tilde{\beta}_k \tilde{u}_{k+1} e_k^T$$

is an Arnoldi decomposition whose starting vector is proportional to $(A - \kappa_1 I) \cdots (A - \kappa_{m-k} I) u_1$.

- Avoid any matrix-vector multiplications in forming the new starting vector.
- Get its Arnoldi decomposition of order k for free.
- For large n the major cost will be in computing UQ.

Chapter 9

Jacobi-Davidson method

References:

(i) Sleijpon, Henk and Van der Vorst, SIAM Matrix Anal, Appl, Vol.17, PP.401-425, 1996

(ii) BIT, Vol.36, PP.595-633, 1996.

(iii) SIAM Sci comp, Vol.20, PP.94-125, 1998

(iv) Lehoucg and Meerbergen, *Using generalized Cauchy Transformation within an inexact rational Krylov subspace method*, 2001

9.1 JOCC(Jacobi Orthogonal Component Correction)

Consider

$$Ax = \lambda x,$$

where A is a nonsymmetric diagonal dominant matrix (i.e., $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$). Let

$$A = \begin{bmatrix} \alpha & C^T \\ b & F \end{bmatrix},$$

with α being the largest diagonal element. Then

$$A \begin{bmatrix} 1 \\ z \end{bmatrix} = \begin{bmatrix} \alpha & C^T \\ b & F \end{bmatrix} \begin{bmatrix} 1 \\ z \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ z \end{bmatrix}. \tag{9.1.1}$$

That is

$$\begin{cases} \lambda = \alpha + C^T z, \\ (F - \lambda I)z = -b. \end{cases} \tag{9.1.2}$$

$EV = \begin{bmatrix} 1 \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ z \end{bmatrix} = e_1 + \begin{bmatrix} 0 \\ z \end{bmatrix}, e_1 \perp \begin{bmatrix} 0 \\ z \end{bmatrix}$ (correction)

Jacobi proposed to solve (9.1.2) by the following Jacobi iteration with $z_1 = 0$

$$\begin{array}{ll} \text{for} & k = 1, 2, \ldots \\ & \theta_k = \alpha + C^T z_k \\ & (D - \theta_k I) z_{k+1} = (D - F) z_k - b \\ \text{end} & \end{array} \tag{9.1.3}$$

where $D = diag(F)$.

Remark: θ_k is not a Ritz value

9.2 Davidson method

Davidson's method as an accelerated JOCC method

Assume $u_k = \begin{bmatrix} 1 \\ z_k \end{bmatrix} \approx EV \equiv x \;\; (Ax = \lambda x)$

and θ_k is the associated EW. The residual

$$r_k = (A - \theta_k I) u_k = \begin{bmatrix} \alpha - \theta_k + c^T z_k \\ (F - \theta_k I) z_k + b \end{bmatrix} \tag{9.2.4}$$

Davidson(1975) proposed computing t_k from

$$(D_A - \theta_k I) t_k = -r_k, \;\; D_A = diag(A) \tag{9.2.5}$$

For the component $\hat{y}_k = \begin{pmatrix} 0 \\ y_k \end{pmatrix}$ of t_k orthog to $u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$(D - \theta_k I) y_k = -(F - \theta_k I) z_k - b = (D - F) z_k - (D - \theta_k I) z_k - b \tag{9.2.6}$$

$$\left((9.2.5) \Rightarrow \left(\begin{bmatrix} \alpha & 0 \\ 0 & D \end{bmatrix} - \theta_k I \right) \begin{pmatrix} * \\ y_k \end{pmatrix} = \begin{bmatrix} * \\ (F - \theta_k I) z_k + b \end{bmatrix} \right)$$

iff

$$(D - \theta_k I)(z_k + y_k) = (D - F). \tag{9.2.7}$$

let $(D - \theta_k I)(z_k + y_k) = (D - \theta_k I) z_{k+1}$

compare (9.2.7), (**??**): $z_k + y_k$ is the z_{k+1} (fixed k)

that we would have obtained with one step

JOCC starting with z_k

Davidson suggested computing Ritz value/vector of A w.r.t

$S_{k+1} =< u_1, \ldots, u_k, \hat{u}_{k+1} >=< u_1, \ldots, u_k, t_k >$
$\overset{\text{O.B.}}{=} < v_1, \ldots, v_{k+1} >$ (orthog basis)

where $u_1 = e_1,\ \hat{u}_{k+1} = u_k + \hat{y}_k = u_k + \begin{pmatrix} 0 \\ y_k \end{pmatrix}$,

$$t_k = \begin{pmatrix} * \\ y_k \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ y_k \end{pmatrix},\ \begin{pmatrix} * \\ 0 \end{pmatrix} = u_1$$

i.e. compute a Ritz pair (θ_{k+1}, u_{k+1}) which is ″nearest″ the target value.
Then compute $r_{k+1} = (A - \theta_{k+1}I)u_{k+1}$ GOTO (9.2.4)

9.3 Jacobi Davidson method

In fact, we want to find the orthog comp for the current approx u_k w.r.t the desired EW u. We are interested in seeing explicitly what happens in the subspace ${u_k}^{\perp}$. Let

$$B = (I - u_k {u_k}^T)A(I - u_k {u_k}^T),\ {u_k}^T u_k = 1$$

$$\Rightarrow A = B + Au_k {u_k}^T + u_k {u_k}^T A - \theta_k u_k {u_k}^T. \tag{9.3.8}$$

where (θ_k, u_k) is a given Ritz pair $\theta_k = \dfrac{{u_k}^T A u_k}{{u_k}^T u_k}$

We are in search of EW λ of A choose to θ_k.

We want to have two correction $v \perp u_k$ s.t.

$$A(u_k + v) = \lambda(u_k + v) \tag{9.3.9}$$

By (9.3.8) and $Bu_k = 0$,

$$\Rightarrow (B - \lambda I)v = -r + (\lambda - \theta_k - {u_k}^T A v)u_k \tag{9.3.10}$$

$$(B - \lambda I)v \in {u_k}^{\perp}$$

Since the left hand side and r have no component in u_k

$$r = Au_k - \theta_k u_k \perp u_k$$

$$\Rightarrow (B - \lambda I)v = -r \tag{9.3.11}$$

$$\underset{v \perp u_k}{\overset{\lambda \approx \theta_k}{\Rightarrow}} (B - \theta_k I)v = -r \tag{9.3.12}$$

$$\Leftrightarrow (I - u_k {u_k}^T)(A - \theta_k I)t = -r,\ t \perp u_k \tag{9.3.13}$$

Remark:

(1) If we take $v = -r \Rightarrow$ Arnoldi or Lanzcos.

(2) If we take $v = -(D_A - \theta_k I)^{-1} r_k \Rightarrow$ Davidson method.

(3) Select sutable approx $\tilde{t} \perp u_k$ for the solution of $(B - \theta_k I)t = -r_k,\ t \perp u_k$

Jacobi-Davidson Algorithm: SIMAX (1996)

1. Start: Choose $v_1(\|v_1\|_2 = 1) w_1 = Av_1, h_{11} = v_1^T w_1$

   Set $V_1 = [v_1], W_1 = [w_1], H_1 = [h_{11}], u = v_1, \theta = h_{11},$

   Compute $r = w_1 - \theta u$

2. Iterate until convergence do

3. Inner loop For $k = 1, \ldots, m-1$ do

   a.Solve (approximately) $t \perp u$

   $(I - uu^T)(A - \theta I)(I - uu^T)t = -r$

   b.Orthog. t against V_k by MGS and $V_{k+1} = [V_k | v_{k+1}], v_{k+1} = t - V_k(V_k^T t)$

   c.$V_{k+1}^T A V_{k+1} \rightarrow H_{k+1}$

   d.Compute the largest (the nearest to the target)

   (θ, s) of $H_{k+1}(\|s\|_2 = 1)$

   e.Compute Ritz vector $u := V_{k+1} s$

   f.Compute residual $r = Au - \theta u$

   g.Test convergence ?

4.Restart set $V_1 = [u], H_1 = [\theta]$ go to 3.

main part: (9.3.12) or (9.3.13)

main step: Solving correction vector t with $t \perp u_k$ from the correction equation

$$\left(I - u_k u_k^T\right) \mathbf{A}(\theta_k)(I - u_k u_k^T)t = -r_k \tag{9.3.14}$$

where $\mathbf{A}(\theta_k) = A - \theta_k I$. There are three methods to solve t.

(a) Method I: Use preconditioning iterative approximations, e.g., GMRES, to solve (9.3.14). The method uses a preconditioner

$$\mathcal{M}_p \equiv \left(I - u_k u_k^T\right) \mathcal{M} \left(I - u_k u_k^T\right) \approx \left(I - u_k u_k^T\right) \mathbf{A}(\theta_k) \left(I - u_k u_k^T\right),$$

where $\mathcal{M}$ is an approximation of $\mathbf{A}(\theta_k)$ and an iterative method to solve Eq. (9.3.14). In each of the iterative steps, it needs to solve the linear system

$$\mathcal{M}_p t = y, \quad t \perp u_k \tag{9.3.15}$$

for a given y. Since $t \perp u_k$, Eq. (9.3.15) can be rewritten as

$$\left(I - u_k u_k^T\right) \mathcal{M} t = y \;\Rightarrow\; \mathcal{M} t = \left(u_k^T \mathcal{M} t\right) u_k + y \equiv \eta_k u_k + y.$$

Hence

$$t = \mathcal{M}^{-1} y + \eta_k \mathcal{M}^{-1} u_k,$$

where

$$\eta_k = -\frac{u_k^T \mathcal{M}^{-1} y}{u_k^T \mathcal{M}^{-1} u_k}.$$

Let $\mathbf{A}(\theta_k) = L + D + U$. Then

$$\mathcal{M} = (D + \omega L) D^{-1} (D + \omega U)$$

is a SSOR preconditioner of $\mathbf{A}(\theta_k)$.
(b) Method II: Since $t \perp u_k$, Eq. (9.3.14) can be rewritten as

$$\mathbf{A}(\theta_k) t = \left(u_k^T \mathbf{A}(\theta_k) t\right) u_k - r_k \equiv \varepsilon u_k - r_k. \tag{9.3.16}$$

Let t_1 and t_2 be approximated solutions of the following linear systems:

$$\mathbf{A}(\theta_k) t = -r_k \quad \text{and} \quad \mathbf{A}(\theta_k) t = u_k,$$

respectively. Then the approximated solution $\tilde{t}$ for (9.3.16) is

$$\tilde{t} = t_1 + \varepsilon t_2 \quad \text{for} \quad \varepsilon = -\frac{u_k^T t_1}{u_k^T t_2}.$$

For the special case, the approximated solution $\tilde{t}$ for (9.3.16) can be

$$\tilde{t} = -\mathcal{M}^{-1} r_k + \varepsilon \mathcal{M}^{-1} u_k \quad \text{for} \quad \varepsilon = \frac{u_k^T \mathcal{M}^{-1} r_k}{u_k^T \mathcal{M}^{-1} u_k},$$

where $\mathcal{M}$ is an approximation of $\mathbf{A}(\theta_k)$.
(c) Method III: Eq. (9.3.16) implies that

$$t = \varepsilon \mathbf{A}(\theta_k)^{-1} u_k - \mathbf{A}(\theta_k)^{-1} r_k = \varepsilon \mathbf{A}(\theta_k)^{-1} u_k - u_k. \tag{9.3.17}$$

Let t_1 be approximated solution of the following linear system:

$$\mathbf{A}(\theta_k) t = u_k.$$

Then the approximated solution $\tilde{t}$ for (9.3.16) is

$$\tilde{t} = \varepsilon t_1 - u_k \quad \text{for} \quad \varepsilon = \left(u_k^T t_1\right)^{-1}.$$

Solve

$$(I - u_k {u_k}^T)(A - \theta_k I)(I - u_k {u_k}^T)t = -r,\ t \perp u_k \tag{9.3.18}$$

$$\text{since } t \perp u_k \Rightarrow \left(I - u_k u_k^T\right) t = t$$

$$\overset{(refeq:23)}{\Rightarrow} \left(I - u_k u_k^T\right)(A - \theta_k I)\, t = -r, t \perp u_k$$

$$\Rightarrow (A - \theta_k I)\, t = \varepsilon u_k - r$$

Determine ε s.t. $t \perp u_k$

$$t = \varepsilon (A - \theta_k I)^{-1} u_k - (A - \theta_k I)^{-1} r$$

Define $t^T u_k = 0$

$$\Rightarrow \varepsilon = \frac{u_k (A - \theta_k I)^{-1} r}{u_k^T (A - \theta_k I)^{-1} u_k}$$

Choose a preconditioner $M \approx A - \theta_k I$

$$\Rightarrow \quad \tilde{t} = \varepsilon M^{-1} u_k - M^{-1} r \tag{9.3.19}$$

$$\text{where } \varepsilon = \frac{u_k^T M^{-1} r}{u_k^T M^{-1} u_k} \tag{9.3.20}$$

Remark:
(1) If we choose $\varepsilon = 0$
$M = D_A - \theta_k I \Rightarrow$ Davidson method
(in this case $\tilde{t} = -M^{-1} r \not\perp u_k$)
(2) If we choose $\varepsilon = 0, M = I$
$\Rightarrow$ Arnoldi or Lanczos
(3) (9.3.17) is equivalent with $t = \varepsilon (A - \theta_k I)^{-1} u_k$, ($\because t \perp u_k$).
Math equation to shift-invert iteration (locally quadratic convergence).
In finite arithmetics, the vector $(A - \theta_k I)^{-1} u_k$ may make a "very small" angle with u_k, s.t. it will be impossible to compute a significant orthogonal search direction.
Discussion :

$A = D_A + E$: strongly diagonally dominant
$\|E\| \ll \|D_A\|$
Assume $\begin{bmatrix} a_{11} & & \varepsilon \\ & \ddots & \\ \varepsilon & & a_{nn} \end{bmatrix}$, "$a_{11}$" : the largest diagonally element.

To Davidson method:

From $r = Au_k - \theta_k u_k = (D_A + E)\, u_k - \theta_k u_k$,

$$\tilde{t} = (D_A - \theta_k I)^{-1} r = u_k + (D_A - \theta_k I)^{-1} E u_k$$
$$(D_A - \theta_k I)^{-1} E u_k \text{ is not small compared with } u_k$$

Davidson method works well for diagonal dominant problem.

$(D_A - \theta_k I)^{-1} E u_k$ not small compared with u_k, $\theta_k \approx a_{11}$,

$\tilde{t}$ is expected to recover some part of significant digits.

To Jacobi-Davidson method:

$$\tilde{t} = \varepsilon\,(D_A - \theta_k I)^{-1} u_k - (D_A - \theta_k I)^{-1} r,\ \tilde{t} \perp u_k$$
$$\varepsilon \text{ is well-determined by } \varepsilon = \frac{u_k^T M^{-1} r}{u_k^T M^{-1} u_k},\ M = (D_A - \theta_k I)\,.$$

$$\begin{aligned} \|\varepsilon\,(D_A - \theta_k I)^{-1} u_k\| &= \|\frac{u_k^T M^{-1} r}{u_k^T M^{-1} u_k}\,(D_A - \theta_k I)^{-1} u_k\| \\ &\le \frac{\|u_k\|\,\|(D_A - \theta_k I)^{-1} r\|}{\|u_k\|\,\|(D_A - \theta_k I)^{-1} u_k\|}\|\,(D_A - \theta_k I)^{-1} u_k\| \\ &= \|\,(D_A - \theta_k I)^{-1} r\|. \end{aligned}$$

$\because u_k \perp r \Rightarrow \{\varepsilon\,(D_A - \theta_k I)^{-1} u_k, (D_A - \theta_k I)^{-1} r\}$ is linearly independent and $\|\varepsilon\,(D_A - \theta_k I)^{-1} u_k\| \approx \|\,(D_A - \theta_k I)^{-1} r\|$. There will be hardly any cacellation in the computation of $\tilde{t}$.

Remark:

$\tilde{t}$ = combination of (Shift-Invert) and (Davidson) ,

where Shift-Invert $= \varepsilon\,(D_A - \theta_k I)^{-1} u_k$, Davidson $= (D_A - \theta_k)^{-1} r$.

Consider $Ax = \lambda x$, λ be simple.

Lemma 9.3.1 *Consider ω with $\omega^T x \neq 0$. Then the map $F_p \equiv \left(I - \frac{x\omega^T}{\omega^T x}\right)(A - \lambda I)\left(I - \frac{x\omega^T}{\omega^T x}\right)$ is a bijection from $\omega^\perp$ to $\omega^\perp$.*

$$\left(\text{Extension : } F_p = \left(I - \frac{uu^T}{u^T u}\right)(A - \theta I)\left(I - \frac{uu^T}{u^T u}\right) t = -r. \quad t \perp u,\ r \perp u,\ t \in u^\perp \underset{F_p}{\to} r \in u^\perp.\right)$$

Proof: Let $y \in \omega^T$ and $F_p y = 0$

Claim : $y = 0$?

$$\begin{aligned} \because F_p y = 0. &\Rightarrow \left(I - \frac{x\omega^T}{\omega^T x}\right)(A - \lambda I)\left(I - \frac{x\omega^T}{\omega^T x}\right) y = 0. \\ &\Rightarrow (A - \lambda I)\, y \parallel x. \\ &\Rightarrow y \text{ and } x \in \mathrm{Ker}(A - \lambda I)^2. \end{aligned}$$

$\because \lambda$ is simple.$\Rightarrow y \parallel x$. But $y \perp \omega$, $x \not\perp \omega \Rightarrow y = 0$

■

Theorem 9.3.2 *Assume that correction equation is solved exactly in each step of JD-algorithm. Choose* $\omega = Au$

$$\underbrace{\left(I - \frac{u\omega^T}{\omega^T u}\right)}_{I-P}(A-\theta I)\left(I - \frac{u\omega^T}{\omega^T u}\right)t = -r, \quad t \perp \omega \tag{9.3.21}$$

Assume $u_k = u \to x$, $\omega_k = \omega \to \omega_k$, $\omega_k^T x \not\to 0$.
Then if u_k *is sufficiently chosen to* x,
then $u_k \to x$ *locally quadratically convergent*

$$\theta_k = \frac{\omega_k^T A u_k}{\omega_k^T u_k} \;\to\; \lambda$$

Proof: $Ax = \lambda x$. Let $x = u + z,\ z \perp \omega$

$$\text{then} \quad (A-\theta I)\, z = -(A-\theta I)\, u + (\lambda - \theta)\, x = -r + (\lambda - \theta)\, x \tag{9.3.22}$$

Consider the exact solution $z_1 \perp \omega$ of (9.3.21)

$$(I-P)(A-\theta I)\, z_1 = -(I-P)\, r \quad (\because (I-P)\, r = r) \tag{9.3.23}$$

$\because \quad x - (u + z_1) = z - z_1$ and $z = x - u$

$$\text{It suffices to show that} \quad \|x - (u+z_1)\| = \|z - z_1\| = O\left(\|z\|^2\right) \tag{9.3.24}$$

Multiplying (9.3.22) by $(I-P)$ and subtracting the result from (9.3.23) yields

$$(I-P)(A-\theta I)(z - z_1) = (\lambda - \theta)(I-P)\, z + (\lambda - \theta)(I-P)\, u \tag{9.3.25}$$

Multiplying (9.3.22) by ω^T and using $r \perp \omega$,

$$\Rightarrow \lambda - \theta = \frac{\omega^T (A - \theta I)\, z}{\omega^T x} \tag{9.3.26}$$

$\because u_k^T x \not\to 0$

$$\Rightarrow \|(\lambda-\theta)(I-P)\, z\| = \|\frac{\omega^T (A-\theta I)\, z}{\omega^T x}(I-P)\, z\| \tag{9.3.27}$$

From (9.3.25), lemma9.3.1 and $(I-P)\, u = 0$

$$\begin{aligned}
\|z - z_1\| &= \|\left[(I-P)(A-\theta_k I)\,|_{\omega_k^\perp}\right]^{-1}(\lambda-\theta)(I-P)\, z\| \\
&= \|\left[(I-P)(A-\theta_k I)\,|_{\omega_k^\perp}\right]^{-1}\frac{\omega_k^T (A-\theta_k I)\, z}{\omega_k^T x}(I-P)\, z\| \\
&= o\left(\|z\|^2\right)
\end{aligned}$$

■

9.3.1 Jacobi Davidson method as on accelerated Newton Scheme

$Ax = \lambda x$, λ : simple.

Choose $\omega^T x = 1 \ \left(\omega^T x \neq 0\right)$

Consider nonlinear equation $F(x) = 0 \Leftrightarrow \quad F(u) = Au - \theta u \ , \ \theta = \theta(u) = \dfrac{\omega^T Au}{\omega^T u}$

$(\|u\| = 1)$ or $\left(\omega^T u = 1\right)$

$$F(u) = \left\{ \begin{array}{c} Au - \theta(u)\, u \\ \text{choose} \quad \theta(u) = \frac{\omega^T Au}{\omega^T u} \\ \omega^T u = 1 \end{array} \right\} = 0$$

$F : \{u | \omega^T u = 1\} \to \omega^{\perp}$

In particular, $r \equiv F(u) = Au - \theta(u)\, u \perp \omega$

If $u_k \approx x$, the next Newton approximation u_{k+1} is given by $u_{k+1} = u_k + t$, where $t \perp \omega$.

$$\left(\because u_{k+1}^T \omega = 1 = (u_k + t)^T \omega = 1 + t^T \omega \Rightarrow \omega^T t = 0\right)$$

$$\left(\frac{\partial F}{\partial u} \mid_{u=u_k}\right) t = F(u_k) = -r$$

$$\left(u_{k+1} = u_k - \left(\frac{\partial F}{\partial u} \mid_{u=u_k}\right)^{-1} F(u_k)\right)$$

The Jacobian of F acts on $\omega^{\perp}$ and is given by

$$\left(\frac{\partial F}{\partial u}|_{u=u_k}\right) t = \left(I - \frac{u_k \omega^T}{\omega^T u_k}\right)(A - \theta_k I)\, t, \ t \perp \omega$$

$$\because \ Au - \theta(u)\, u = Au - \frac{\omega^T Au}{\omega^T u} u$$

$$\begin{aligned} \frac{\partial F}{\partial u} &= A - \theta(u)\, I - \frac{-\left(\omega^T Au\right) u\omega^T + \left(\omega^T u\right) u\omega^T A}{\left(w^T u\right)^2} \\ &= A - \theta I + \frac{\omega^T Au}{\left(\omega^T u\right)^2} u\omega^T - \frac{u\omega^T A}{\omega^T u} \end{aligned}$$

On the other hand,

$$\left(I - \frac{u\omega^T}{\omega^T u}\right)(A - \theta I) = A - \theta I - \frac{u\omega^T A}{\omega^T u} + \frac{\omega^T Au}{\left(\omega^T u\right)^2} u\omega^T$$

Hence the correction equation of Newton method read as :

$$\begin{array}{ll} t \perp \omega, & \left(I - \dfrac{u_k \omega^T}{\omega^T u_k}\right)(A - \theta_k I)\, t = -r \\ \Longleftrightarrow & \text{correction equation of JD} \end{array}$$

9.3.2 Jacobi-Davidson with harmonic Ritz values

Ritz values : $V_k \subseteq \mathbb{C}^n$
(θ_k, u_k) is a Ritz pair.

$$Au_k - \theta_k u_k \perp V_k \tag{9.3.28}$$

Harmonic Ritz values : (Inverse of A implicitly)
$\theta_k \in \mathbb{C}$ is a harmonic Ritz value of A with respect to W_k, if θ_k^{-1} is a Ritz value of A^{-1} with respect to W_k
To avoid computing A^{-1}
Remark: A is a normal matrix. A^{-1} is normal.

Theorem 9.3.3 *Let $V_k =< v_1, v_2, \ldots, v_k >$, $\theta_k \in \mathbb{C}$ is a harmonic Ritz value of A with respect to $W_k = AV_k$*

$$\iff \quad Au_k - \theta_k u_k \perp AV_k \quad \textit{for some } u_k \in V_k \tag{9.3.29}$$

If $AV_k = W_k =< w_1, \ldots, w_k >$ and

$$H_k = \left(W_k^T V_k^T\right)^{-1} \left(W_k^T A V_k\right) \tag{9.3.30}$$

then (9.3.29) $\iff H_k S = \theta_k S$, $u_k = V_k S$

Remark: Compute $H_k := \left(V_k^T A^T V_k\right)^{-1} \left(V_k^T A^T A V_k\right)$. Compute $H_k S = \theta_k S$.
Proof: By (9.3.28). $\left(\theta_k^{-1}, Au_k\right)$ is a Ritz pair of A^{-1} with respect to $W_k = AV_k$, some $u_k \in V_k$.

$$\begin{aligned}
&\underbrace{A^{-1}\left(Au_k\right) - \theta_k^{-1}\left(Au_k\right)} \perp AV_k \\
&\quad = -\theta_k^{-1}\left(Au_k - \theta_k u_k\right) \perp AV_k \\
\iff &\quad Au_k - \theta_k u_k \perp AV_k \\
\iff &\quad (9.3.29) \\
\iff &\quad W_k^T\left(Au_k - \theta_k u_k\right) = 0 \\
\iff &\quad W_k^T\left(AV_k S - \theta_k V_k S\right) = 0 \\
\iff &\quad \left(W_k^T A V_k\right) S = \theta_k \left(W_k^T V_k\right) S
\end{aligned}$$

■

Remark: If $V_k \in \mathbb{R}^{n\times n}$ (In general, $V_k \in \mathbb{R}^{n\times k}$)

$$\begin{aligned}
H_k &:= \left(W_k^T V_k\right)^{-1}\left(W_k^T A V_k\right) \\
&= V_k^{-1} W_k^{-T} W_k^T A V_k \\
H_k^{-1} &\approx A^{-1}
\end{aligned}$$

Bi-orthogonalization basis construction
$V_k =< v_1, \ldots, v_k >$, $W_k =< w_1, \ldots, w_k >$, $W_k = AV_k$ and $L_k = W_k^T V_k$(Lower triangular), we say V_k and W_k be bi-orthogonal.
Let $H_k = \left(W_k^T V_k\right)^{-1}\left(W_k^T A V_k\right)$

If (θ_k, S) is eigenpair of $H_k \Longleftrightarrow (\theta_k, V_k S)$ is a harmonic Ritz pair.
We bi-orthogonalize t with respect to V_k and W_k.
$\tilde{t} := t - V_k L_k^{-1} W_k^T t$, $v_k = \frac{\tilde{t}}{\|\tilde{t}\|_2}$
$V_k = \{v\}$, $W_k = \{w\}$.
$V_{k+1} =< V_k | v_{k+1} >$, $W_{k+1} =< W_k | w_{k+1} >$.
Correction equation:

$$\begin{cases} \left(I - \frac{u_k w_k^T}{w_k^T u_k}\right)(A - \theta_k I)\left(I - \frac{u_k w_k^T}{w_k^T u_k}\right) t = -r \\ t \perp w_k = Au_k, \quad \text{where} \quad H_k S = \theta_k S, \quad u_k = V_k S \end{cases}$$

Solve correction equation:
$\Longleftrightarrow (A - \theta_k I)\, t = \varepsilon u_k - r$
$\Longleftrightarrow t = \varepsilon (A - \theta_k I)^{-1} u_k - (A - \theta_k I)^{-1} r$
$0 = w_k^T t = \varepsilon w_k^T (A - \theta_k I)^{-1} u_k - w_k^T (A - \theta_k I)^{-1} r$
$M \approx A - \theta_k I$ preconditioner , $\varepsilon = \frac{w_k^T M^{-1} r}{w_k^T M^{-1} u_k}$

Remark:

$$\begin{cases} G_k &= Q_k^T A Q_k \\ H_k &= \left(V_k^T A V_k\right)^{-1} \left(V_k^T A^T A V_k\right) \end{cases}$$
$$H_k^{-1} = \left(V_k^T A^T A V_k\right)^{-1} \left(V_k^T A V_k\right)$$
$$G_k^{-1} \neq H_k^{-1}$$

Algorithm 1 : JD with Ritz value and orthogonal
Algorithm 2 : JD with harmonic Ritz value and bi-orthogonal
(1) Start :
choose v_1 ($\|v_1\|_2 = 1$), $w_1 = Av_1$
$l_{11} = w_1^T v_1$, $h_{11} = w_1^T w_1$
$l = 1$, $V_1 = [v_1]$, $W_1 = [w_1]$, $L_1 = [l_{11}]$, $H_1 = [h_{11}]$
$u = v_1$, $w = w_1$, $\theta = \frac{h_{11}}{l_{11}}$. Compute $r = Au - \theta u$.
(2) Iterate until convergence do :
(3) Inner loop. For $k = l_1, \ldots, m-1$ do
Solve approximation $t \perp w$

$$\left(I - \frac{uw^T}{w^T u}\right)(A - \theta I)\left(I - \frac{uw^T}{w^T u}\right) t = -r$$
$$\left(I - \frac{uu^T}{u^T u}\right)(A - \theta I)\left(I - \frac{uu^T}{u^T u}\right) t = -r$$

To solve ε
• Bi-orthogonalize t against V_k and W_k.

$$\tilde{t} = t - V_k L_k^{-1} W_k^T t, \quad v_{k+1} = \frac{\tilde{t}}{\|\tilde{t}\|_2}$$
$$\tilde{t} = t - V_k \left(V_k^T V_k\right)^{-1} V_k t$$

• Compute

$$w_{k+1} = Av_{k+1}$$
$$W_{k+1} = [W_k | w_{k+1}]$$
$$V_{k+1} = [V_k | v_{k+1}] .$$

• Compute

$$L_{k+1} = W_{k+1}^T V_{k+1}$$
$$H_{k+1} = L_{k+1}^{-1} \left(W_{k+1}^T V_{k+1} \right)$$
$$= \left(V_{k+1}^T A V_{k+1} \right)^{-1} \left(V_{k+1}^T A^T A V_{k+1} \right)$$
$$H_{k+1} = V_k^T A V_k$$

• Compute the smallest, largest eigenpair (θ_k, s) of H_{k+1}.
• Compute the harmonic Ritz vector $u_i = \frac{V_{k+1} S}{\|V_{k+1} S\|}$
• Compute $w := Au$.
• Compute $r := Au - \theta u \, (= w - \theta u)$.
• Test convergence ?
STOP if no go to Inner loop.
(4) Restart.

9.4 Jacobi-Davidson Type method for Generalized Eigenproblems

Generalized eigenvalue problem

(1) $$\mu Ax = \lambda Bx \qquad (|\mu|^2 + |\lambda|^2 = 1)$$

$((\mu, \lambda), x)$ eigenpair.

The updating process for approx. EV:

Let (θ, u) be Ritz pair of (A, B). Then

$$r \equiv Au - \theta Bu \perp u.$$

The goal is to find an update z for u:

(2) $$z \perp u \;\text{ and }\; A(u + z) = \lambda B(u + z).$$

$\Longrightarrow$

(3) $$\begin{cases} \lambda = \frac{u^* A(u+z)}{u^* B(u+z)} \\ z \perp u \;\text{ and }\; (I - \frac{uu^*}{u^*u})(A - \lambda B)(I - \frac{uu^*}{u^*u})z \\ \qquad\qquad = -r \equiv -(Au - \lambda Bu) \end{cases}$$

In practice, $\theta = \frac{u^* Au}{u^* Bu}$, $z \perp u$ and

(4) $$(I - \frac{uu^*}{u^*u})(A - \theta B) \mid_{u^\perp} z = -r \equiv -(Au - \theta Bu)$$

Other projection for approx. EV:

Assume

$$r \equiv Au - \theta Bu \perp w \quad \text{for some } w.$$

We look for an update z of u which is orthog. to $\tilde{u}$ ($\tilde{u} \perp u$). i.e.

$$z \perp \tilde{u} \ \text{ and } \ A(u+z) = \lambda B(u+z). \tag{5}$$

For convenience, $\tilde{u} = Bu$. Similarly, select $\tilde{w} \perp w$ and consider

$$P = \frac{\tilde{w}w^*}{w^*\tilde{w}} \ \text{ and } \ Q = \frac{u\tilde{u}^*}{\tilde{u}^*u}$$

$\Longrightarrow$

$$(A-\lambda B)x = 0 \Longleftrightarrow \begin{cases} P(A-\lambda B)(Qx + (I-Q)x) = 0 \ \ \& \\ (I-P)(A-\lambda B)(Qx+(I-Q)x) = 0. \end{cases} \tag{6}$$

With

$$\alpha \equiv \frac{w^*Au}{w^*u},\ \beta = \frac{w^*Bu}{w^*u},\ a = Au - \alpha u,\ b = Bu - \beta u \tag{7}$$

and

$$u' \equiv \left(I - \frac{\tilde{w}w^*}{w^*\tilde{w}}\right) u.$$

(5) is equiv. to

$$\begin{cases} \lambda = \frac{w^*A(u+z)}{w^*B(u+z)} \\ z \perp \tilde{u} \ \ \& \ \ (I - \frac{\tilde{w}w^*}{w^*\tilde{w}})(A-\lambda B)_{\tilde{u}^\perp} z \\ \qquad\qquad = -(a-\lambda b) - (\alpha - \lambda\beta)u'. \end{cases} \tag{8}$$

In practice, let

$$\theta = \frac{\alpha}{\beta} \ \text{ and } \ r = a - \theta b = Au - \theta Bu.$$

Lemma 9.4.1 *Let $(A-\lambda B)x = 0$. Consider w and $\tilde{u}$ with $\tilde{u}^*x \neq 0$ and $(Bx)^*w \neq 0$. Then the map*

$$F_p \equiv (I - \frac{Bxw^*}{w^*Bx})(A-\lambda B)(I - \frac{x\tilde{u}^*}{\tilde{u}^*x})$$

is a bijection from $\tilde{u}^\perp$ onto $w^\perp$.

Proof: Suppose $y \perp \tilde{u}$ and $F_p y = 0 \Longrightarrow y = 0$.

Theorem 9.4.1 *Choose $\tilde{w} = Bu$. Assume $\tilde{u}$ and w conv. and $\tilde{u}^*x$ and $w^*Bx \nrightarrow 0$. Then if the initial $u \approx x$, the seq. of u conv. to x quadratically and $\theta = w^*Au/w^*Bu \longrightarrow \lambda$.*

Proof: Suppose $(A-\lambda B)x = 0$, with $x = u + z$ for $z \perp \tilde{u}$. Then

$$(A-\theta B)z = -(A-\theta B)u + (\lambda - \theta)Bx = -r + (\lambda - \theta)Bx. \tag{10}$$

Solve

$$(I-P)(A-\theta B)\,|_{\tilde{u}^\perp}\, z_1 = -(I-P)r. \tag{11}$$

$\because \quad x-(u+z_1) = z-z_1$ and $z = x-u$. It suffices to show

$$\|x-(u+z_1)\| = \|z-z_1\| = O(\|z\|^2).$$

$(I-P)\times(10)-(11) \Longrightarrow$

$$\begin{aligned}(I-P)(A-\theta B)(z-z_1) &= (\lambda-\theta)(I-P)Bz \\ &+ (\lambda-\theta)(I-P)Bu.\end{aligned}$$

$w^*\times(10)$ and using $r\perp w \Longrightarrow$

$$\lambda-\theta = \frac{w^*(A-\theta B)z}{w^*Bx}. \tag{12}$$

By assumption and (12) $\Longrightarrow$

$$\begin{aligned}\|(\lambda-\theta)(I-P)Bz\| &= \|\frac{w^*(A-\theta B)z}{w^*Bx}(I-P)Bz\| \\ &= O(\|z\|^2),\end{aligned}$$

provided $(I-P)(A-\theta B)|_{\tilde{u}^\perp}$ nonsing. (by Lemma 1) and $(I-P)Bu=0$. ($\because\ \tilde{w}=Bu$)

In practice, $w=\tilde{w}=Bu$, $\tilde{u}=B^*w$.

Equiv. formulation for the correction:

correction:

$$(I-\frac{\tilde{w}w^*}{w^*\tilde{w}})(A-\theta B)|_{\tilde{u}^\perp} z_1 = -r, \quad z_1\perp\tilde{u}. \tag{13}$$

is equiv. to

$$\begin{bmatrix} A-\theta B & \tilde{w} \\ \tilde{u}^* & 0 \end{bmatrix}\begin{bmatrix} z \\ \epsilon \end{bmatrix} = \begin{bmatrix} -r \\ 0 \end{bmatrix}, \tag{14}$$

where $\epsilon = -w^*(A-\theta B)z/w^*\tilde{w}$.

Theorem 9.4.2 *The solution (13) is given by*

$$z = (A-\theta B)^{-1}(-r+\tilde{w}) = -u+\epsilon(A-\theta B)^{-1}\tilde{w} \tag{15}$$

with $\epsilon = \dfrac{\tilde{u}^*u}{\tilde{u}^*(A-\theta B)^{-1}\tilde{w}}$.

Proof: With z in (15) $\Rightarrow \tilde{u}\perp z$, and $(A-\theta B)z = -r+\epsilon\tilde{w}$. Since $r\perp w \Rightarrow$ (13) holds.

Bibliography

[1] R. Barrett, M. Berry, T. F. Chan, and J. Demmel, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems*, SIAM, Philadelphia, 2000.

[3] B. N. Datta, *Numerical linear algebra and applications*, Pacific Grove :Brooks/Cole Pub., 1995.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations, 3rd ed*, The Johns Hopkins University Press, 1996.

[5] L. A. Hageman, D. M. Young, *Applied iterative methods*, New York, Academic Press, 1981.

[6] N. J. Higham, *Accuracy and stability of numerical algorithms, 2nd ed.*, SIAM, , Philadelphia, 2002.

[7] G. W. Stewart, *Introduction to matrix computations*, New York, Academic Press, 1973.

[8] G. W. Stewart, *Matrix algorithms*, SIAM, Philadelphia, 1998.

[9] J. H. Wilkinson, *The Algebraic Eigenvalue Problems*, Oxford Science Publications, 1965.

Index