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 $A \in \mathbb{R}^{n\times n}$ is large, sparse and symmetric. There exists an orthogonal matrix Q, which transforms A to a tridiagonal matrix T.

$$Q^T AQ = T \equiv \text{ tridiagonal.} \tag{7.1.1}$$

Remark 7.1.1 **(a)** *Such Q can be generated by Householder transformations or Givens rotations.*

(b) *Almost for all A (i.e. all eigenvalues are distinct) and almost for any $q_1 \in \mathbb{R}^n$ with $\|q_1\|_2 = 1$, there exists an orthogonal matrix Q with first column q_1 satisfying (7.1.1). q_1 determines T uniquely up to the sign of the columns (that is, we can multiply each column with -1).*

Let $(x \in \mathbb{R}^n)$

$$K[x, A, m] = [x, Ax, A^2x, \cdots, A^{m-1}x] \in \mathbb{R}^{n\times m}. \tag{7.1.2}$$

$K[x, A, m]$ is called a Krylov-matrix. Let

$$\mathcal{K}(x, A, m) = Range(K[x, A, m]) = Span(x, Ax, \cdots, A^{m-1}x). \tag{7.1.3}$$

$\mathcal{K}(x, A, m)$ is called the Krylov-subspace generated by $K[x, A, m]$.

Remark 7.1.2 *For each $H \in \mathbb{C}^{n\times m}$ or $\mathbb{R}^{n\times m}$ $(m \le n)$ with $rank(H) = m$, there exists an $Q \in \mathbb{C}^{n\times m}$ or $\mathbb{R}^{n\times m}$ and an upper triangular $R \in \mathbb{C}^{m\times m}$ or $\mathbb{R}^{m\times m}$ with $Q^*Q = I_m$ such that*

$$H = QR. \tag{7.1.4}$$

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*

$$K[x, A, m] = Q_m R_m \tag{7.1.5}$$

is an QR factorization, then $Q_m^* A Q_m = T_m$ is an $m \times m$ tridiagonal matrix and satisfies

$$AQ_m = Q_m T_m + r_m e_m^T, \quad Q_m^* r_m = 0. \tag{7.1.6}$$

(b) *Let $\|x\|_2 = 1$. If $Q_m \in \mathbb{C}^{n\times m}$ 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

$$K[x, A, m] = [x, Ax, \cdots, A^{m-1}x] = Q_m[e_1, T_m e_1, \cdots, T_m^{m-1} e_1] \tag{7.1.7}$$

is an QR factorization of $K[x, A, m]$.

Proof: (a) Since

$$A\mathcal{K}(x, A, j) \subset \mathcal{K}(x, A, j+1), \quad j < m. \tag{7.1.8}$$

From (7.1.5), we have

$$Span(q_1, \cdots, q_i) = \mathcal{K}(x, A, i), \quad i \le m. \tag{7.1.9}$$

So we have

$$q_{i+1} \perp \mathcal{K}(x, A, i) \overset{(7.1.8)}{\supset} A\mathcal{K}(x, A, i-1) = A(span(q_1, \cdots, q_{i-1})).$$

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 (7.1.6). 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

$$AQ_mR_m = Q_mR_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} + Q_mQ_m^*A^mxe_m^T + (I - Q_mQ_m^*)A^mxe_m^T.$$

Then

$$\begin{aligned} AQ_m &= Q_m[R_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} + Q_m^*A^mxe_m^T]R_m^{-1} + (I - Q_mQ_m^*)A^mxe_m^TR_m^{-1} \\ &= Q_m[R_m \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} R_m^{-1} + \gamma Q_m^*A^mxe_m^T] + \underbrace{\gamma(I - Q_mQ_m^*)A^mx}_{r_m} e_m^T \\ &= Q_mH_m + r_me_m^T \text{ with } Q_m^*r_m = 0, \end{aligned}$$

where H_m is an upper Hessenberg matrix. But $Q_m^*AQ_m = H_m$ is Hermitian, so $H_m = T_m$ is tridiagonal.

(b) We check (7.1.7):

$x = Q_me_1$ coincides the first column. Suppose that i-th columns are equal, i.e.

$$\begin{aligned} A^{i-1}x &= Q_mT_m^{i-1}e_1 \\ A^ix &= AQ_mT_m^{i-1}e_1 \\ &= (Q_mT_m + r_me_m^T)T_m^{i-1}e_1 \\ &= Q_mT_m^ie_1 + r_me_m^TT_m^{i-1}e_1. \end{aligned}$$

But $e_m^TT_m^{i-1}e_1 = 0$ for $i < m$. Therefore, $A^ix = Q_mT^ie_1$ the $(i+1)$-th columns are equal. It is clearly that $(e_1, T_me_1, \cdots, T_m^{m-1}e_1)$ is an upper triangular matrix. ■

Theorem 7.1.1 *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 $Q^*AQ = T$, $\tilde{Q}^*A\tilde{Q} = \tilde{T}$ and $Q_1e_1 = \tilde{Q}e_1$

$$\begin{aligned} &\Rightarrow \quad K[q_1, A, n] = QR = \tilde{Q}\tilde{R} \\ &\Rightarrow \quad Q = \tilde{Q}D, \quad R = D\tilde{R}. \end{aligned}$$

Substitute Q by QD, where $D = diag(\epsilon_1, \cdots, \epsilon_n)$ with $|\epsilon_i| = 1$. Then

$$(QD)^*A(QD) = D^*Q^*AQD = D^*TD = \text{ tridiagonal.}$$

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

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

$$AQ = QT, \tag{7.1.10}$$

where

$$Q = [q_1, \cdots, q_n] \text{ and } T = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

It implies that the j-th column of (7.1.10) forms:

$$Aq_j = \beta_{j-1}q_{j-1} + \alpha_j q_j + \beta_j q_{j+1}, \tag{7.1.11}$$

for $j = 1, \cdots, n$ with $\beta_0 = \beta_n = 0$. By multiplying (7.1.11) by q_j^T we obtain

$$q_j^T Aq_j = \alpha_j. \tag{7.1.12}$$

Define $r_j = (A - \alpha_j I)q_j - \beta_{j-1}q_{j-1}$. Then

$$r_j = \beta_j q_{j+1}$$

with

$$\beta_j = \pm\|r_j\|_2 \tag{7.1.13}$$

and if $\beta_j \neq 0$ then

$$q_{j+1} = r_j/\beta_j. \tag{7.1.14}$$

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:

$$\begin{array}{l} j = 0,\ r_0 = q_1\ ,\ \beta_0 = 1\ ,\ q_0 = 0 \\ \text{Do}\quad \text{while } (\beta_j \neq 0) \\ \qquad q_{j+1} = r_j/\beta_j\ ,\ j := j + 1 \\ \qquad \alpha_j = q_j^T Aq_j\ , \\ \qquad r_j = (A - \alpha_j I)q_j - \beta_{j-1}q_{j-1}, \\ \qquad \beta_j = \|r_j\|_2. \end{array} \tag{7.1.15}$$

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 $A \in \mathbb{R}^{n\times n}$ and $w \in \mathbb{R}^n$ 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.

$v_i := 0 \ \ (i = 1, \cdots, n)$
$\beta_0 := 1$
$j := 0$
Do while $(\beta_j \neq 0)$
  if $(j \neq 0)$, then
    for $i = 1, \cdots, n$,
      $t := w_i, w_i := v_i/\beta_j, v_i := -\beta_j t.$
    end for
  end if
  $v := Aw + v,$
  $j := j + 1,$
  $\alpha_j := w^T v,$
  $v := v - \alpha_j w,$
  $\beta_j := \|v\|_2.$

Remark 7.1.3 **(a)** *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.*

(b) *The iteration stops before complete tridiagonalizaton if q_1 is contained in a proper invariant subspace. From the iteration (7.1.15) we have*

$$A(q_1, \cdots, q_m) = (q_1, \cdots, q_m) \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \beta_{m-1} \\ & \ddots & \ddots & \\ & & \beta_{m-1} & \alpha_m \end{bmatrix} + \underbrace{(0, \cdots, 0, \overbrace{\beta_m q_{m+1}}^{r_m})}_{r_m e_m^T}$$

$$\beta_m = 0 \ \ \textit{if and only if} \ \ r_m = 0.$$

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 7.1.2 *Let A be symmetric and q_1 be a given vector with $\|q_1\|_2 = 1$. The Lanczos iterations* (7.1.15) *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*

$$AQ_j = Q_j T_j + r_j e_j^T \tag{7.1.16}$$

with

$$T_j = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ & & \beta_{j-1} & \alpha_j \end{bmatrix} \ \textit{and} \ \ Q_j = [q_1, \cdots, q_j]$$

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 (7.1.15) that (7.1.16) 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 rarely in practice. Consequently, other explanations for the convergence of $T_j's$ eigenvalues must be sought.

Theorem 7.1.3 *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 $Y_j \in \mathbb{R}^{n \times j}$ 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 (7.1.16) 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 7.1.4 *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 7.1.4 *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 7.1.5 *Using the same notation as Theorem 7.1.4*

$$\lambda_n \leq \theta_j \leq \lambda_n + \frac{(\lambda_1 - \lambda_n)\tan^2(\phi_n)}{c_{j-1}^2(1 + 2\rho_n)},$$

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 7.1.4 with A replaced by $-A$. ■

Example 7.1.1

$$\begin{aligned} L_{j-1} &\equiv \frac{1}{[C_{j-1}(2\frac{\lambda_1}{\lambda_2} - 1)]^2} \geq \frac{1}{[C_{j-1}(1 + 2\rho_1)]^2} \\ R_{j-1} &= (\frac{\lambda_2}{\lambda_1})^{2(j-1)} \quad \textit{power method} \end{aligned}$$

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.

7.1.1 Reorthogonalization

Since

$$AQ_j = Q_j T_j + r_j e_j^T,$$

let

$$\begin{aligned} AQ_j - Q_j T_j &= r_j e_j^T + F_j && (7.1.17)\\ I - Q_j^T Q_j &= C_j^T + \Delta_j + C_j, && (7.1.18) \end{aligned}$$

where C_j is strictly upper triangular and Δ_j is diagonal. (For simplicity, suppose $(C_j)_{i,i+1} = 0$ and $\Delta_i = 0$.)

Definition 7.1.1 *θ_i and $y_i \equiv Q_j s_i$ are called Ritz value and Ritz vector, respectively, if $T_j s_i = \theta_i s_i$.*

Let $\Theta_j \equiv \mathrm{diag}(\theta_1, \cdots, \theta_j) = S_j^T T_j S_j$ where $S_j = \left[\begin{array}{ccc} s_1 & \cdots & s_j \end{array} \right]$.

Theorem 7.1.6 (Paige Theorem) *Assume that* (a) *S_j and Θ_j are exact ! (Since $j \ll n$).* (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*

$$\begin{aligned} F_j^T Q_j - Q_j^T F_j &= K_j - K_j^T, \\ \Delta_j T_j - T_j \Delta_j &\equiv N_j - N_j^T, \\ G_j &= S_j^T (K_j + N_j) S_j \equiv (r_{ik}). \end{aligned}$$

Then

(a) *$y_i^T q_{j+1} = r_{ii}/\beta_{ji}$, where $y_i = Q_j s_i$, $\beta_{ji} = \beta_j s_{ji}$.*

(b) *For $i \neq k$,*

$$(\theta_i - \theta_k) y_i^T y_k = r_{ii}(\frac{s_{jk}}{s_{ji}}) - r_{kk}(\frac{s_{ji}}{s_{jk}}) - (r_{ik} - r_{ki}). \qquad (7.1.19)$$

Proof: Multiplied (7.1.17) from left by Q_j^T, we get

$$Q_j^T A Q_j - Q_j^T Q_j T_j = Q_j^T r_j e_j^T + Q_j^T F_j, \qquad (7.1.20)$$

which implies that

$$Q_j^T A^T Q_j - T_j Q_j^T Q_j = e_j r_j^T Q_j + F_j^T Q_j. \qquad (7.1.21)$$

Subtracted (7.1.20) from (7.1.21), we have

$$\begin{aligned} & (Q_j^T \gamma_j) e_j^T - e_j (Q_j^T \gamma_j)^T \\ = \; & (C_j^T T_j - T_j C_j^T) + (C_j T_j - T_j C_j) + (\Delta_j T_j - T_j \Delta_j) + F_j^T Q_j - Q_j F_j^T \\ = \; & (C_j^T T_j - T_j C_j^T) + (C_j T_j - T_j C_j) + (N_j - N_j^T) + (K_j - K_j^T). \end{aligned}$$

This implies that

$$(Q_j^T r_j)e_j^T = C_j T_j - T_j C_j + N_j + K_j.$$

Thus,

$$\begin{aligned} y_i^T q_{j+1}\beta_{ji} = s_i^T(Q_j^T r_j)e_j^T s_i &= s_i^T(C_j T_j - T_j C_j)s_i + s_i^T(N_j + K_j)s_i \\ &= (s_i^T C_j s_i)\theta_i - \theta_i(s_i^T C_j s_i) + r_{ii}, \end{aligned}$$

which implies that

$$y_i^T q_{j+1} = \frac{r_{ii}}{\beta_{ji}}.$$

Similarly, (7.1.19) can be obtained by multiplying (7.1.20) from left and right by s_i^T and s_i, respectively. ■

Remark 7.1.5 *Since*

$$y_i^T q_{j+1} = \frac{r_{ii}}{\beta_{ji}} = \begin{cases} O(esp), & \text{if } |\beta_{ji}| = O(1), \text{ (not converge!)} \\ O(1), & \text{if } |\beta_{ji}| = O(esp), \text{ (converge for } (\theta_j, y_j)) \end{cases}$$

we have that $q_{j+1}^T y_i = O(1)$ when the algorithm converges, i.e., q_{j+1} is not orthogonal to $< Q_j >$ where $Q_j s_i = y_i$.

(i) Full Reorthogonalization by MGS:

Orthogonalize q_{j+1} to all $q_1, \cdots, q_j$ by

$$q_{j+1} := q_{j+1} - \sum_{i=1}^{j}(q_{j+1}^T q_i)q_i.$$

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 $P_0 = I - 2v_0 v_0^T / v_0^T v_0$ so that $P_0 r_0 = e_1$;
$\alpha_1 := q_1^T A q_1$;
Do $j = 1, \cdots, n-1$,
  $r_j := (A - \alpha_j)q_j - \beta_{j-1}q_{j-1} \quad (\beta_0 q_0 \equiv 0)$,
  $w := (P_{j-1} \cdots P_0)r_j$,
  Determine $P_j = I - 2v_j v_j^T / v_j^T v_j$ such that $P_j w = (w_1, \cdots, w_j, \beta_j, 0, \cdots, 0)^T$,
  $q_{j+1} := (P_0 \cdots P_j)e_{j+1}$,
  $\alpha_{j+1} := q_{j+1}^T A q_{j+1}$.

This is the complete reorthogonalization Lanczos scheme.

(ii) 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

(iii) Restart after m-steps

(Do full Reorthogonalization)

(iv) Partial Reorthogonalization

Do reorthogonalization with previous (e.g. $k = 5$) Lanczos vectors $\{q_1, \ldots, q_k\}$

For details see the books:

Parlett: "Symmetric Eigenvalue problem" (1980) pp.257–

Golub & Van Loan: "Matrix computation" (1981) pp.332–

To (7.1.19): 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 answer is using **the implicit Restart Lanczos algorithm:**

Let

$$AQ_j = Q_j T_j + r_j e_j^T$$

be a Lanczos decomposition.

- In principle, we can keep expanding the Lanczos decomposition until the Ritz pairs have converged.
- Unfortunately, it is limited by the amount of memory to storage of Q_j.
- Restarted the Lanczos process once j becomes so large that we cannot store Q_j.
  - Implicitly restarting method
- 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.

7.1.2 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

$$p(t) = (t - \mu_{k+1}) \cdots (t - \mu_m).$$

7.1.3 Implicitly restarted algorithm

Let

$$AQ_m = Q_m T_m + \beta_m q_{m+1} e_m^T \tag{7.1.22}$$

be a Lanczos 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

$$A\tilde{Q}_k = \tilde{Q}_k \tilde{T}_k + \tilde{\beta}_k \tilde{q}_{k+1} e_k^T$$

of order k with starting vector $p(A)u_1$.

Let $\nu_1, \ldots, \nu_m$ be eigenvalues of T_m and suppose that $\nu_1, \ldots, \nu_{m-k}$ correspond to the part of the spectrum we are not interested in. Then take

$$p(t) = (t - \nu_1)(t - \nu_2) \cdots (t - \nu_{m-k}).$$

The starting vector $p(A)u_1$ is equal to

$$\begin{aligned} p(A)u_1 &= (A - \nu_{m-k}I) \cdots (A - \nu_2 I)(A - \nu_1 I)u_1 \\ &= (A - \nu_{m-k}I) \left[\cdots [(A - \nu_2 I) \left[(A - \nu_1 I)u_1 \right]] \right]. \end{aligned}$$

In the first, we construct a Lanczos decomposition with starting vector $(A - \nu_1 I)u_1$. From (7.1.22), we have

$$\begin{aligned} (A - \nu_1 I)Q_m &= Q_m(T_m - \nu_1 I) + \beta_m q_{m+1} e_m^T \\ &= Q_m U_1 R_1 + \beta_m q_{m+1} e_m^T, \end{aligned} \tag{7.1.23}$$

where

$$T_m - \nu_1 I = U_1 R_1$$

is the QR factorization of $T_m - \kappa_1 I$. Postmultiplying by U_1, we get

$$(A - \nu_1 I)(Q_m U_1) = (Q_m U_1)(R_1 U_1) + \beta_m q_{m+1}(e_m^T U_1).$$

It implies that

$$AQ_m^{(1)} = Q_m^{(1)} T_m^{(1)} + \beta_m q_{m+1} b_{m+1}^{(1)T},$$

where

$$Q_m^{(1)} = Q_m U_1, \quad T_m^{(1)} = R_1 U_1 + \nu_1 I, \quad b_{m+1}^{(1)T} = e_m^T U_1.$$

($Q_m^{(1)}$: one step of single shifted QR algorithm)

Remark 7.1.6

- $Q_m^{(1)}$ *is orthonormal.*
- *By the definition of* $T_m^{(1)}$*, we get*

$$U_1 T_m^{(1)} U_1^T = U_1(R_1 U_1 + \nu_1 I)U_1^T = U_1 R_1 + \nu_1 I = T_m. \tag{7.1.24}$$

  Therefore, $\nu_1, \nu_2, \ldots, \nu_m$ *are also eigenvalues of* $T_m^{(1)}$*.*

- *Since* T_m *is tridiagonal and* U_1 *is the Q-factor of the QR factorization of* $T_m - \nu_1 I$*, it implies that* U_1 *and* $T_m^{(1)}$ *are upper Hessenberg. From* (7.1.24)*,* $T_m^{(1)}$ *is symmetric. Therefore,* $T_m^{(1)}$ *is also tridiagonal.*
- *The vector* $b_{m+1}^{(1)T} = e_m^T U_1$ *has the form*

$$b_{m+1}^{(1)T} = \begin{bmatrix} 0 & \cdots & 0 & U_{m-1,m}^{(1)} & U_{m,m}^{(1)} \end{bmatrix};$$

  i.e., only the last two components of $b_{m+1}^{(1)}$ *are nonzero.*

- *For on postmultiplying (7.1.23) by* e_1*, we get*

$$(A - \nu_1 I)q_1 = (A - \nu_1 I)(Q_m e_1) = Q_m^{(1)} R_1 e_1 = r_{11}^{(1)} q_1^{(1)}.$$

  Since T_m *is unreduced,* $r_{11}^{(1)}$ *is nonzero. Therefore, the first column of* $Q_m^{(1)}$ *is a multiple of* $(A - \kappa_1 I)q_1$*.*

Repeating this process with $\nu_2, \ldots, \nu_{m-k}$, the result will be a Krylov decomposition

$$AQ_m^{(m-k)} = Q_m^{(m-k)} T_m^{(m-k)} + \beta_m q_{m+1} b_{m+1}^{(m-k)T}$$

with the following properties

i. $Q_m^{(m-k)}$ is orthonormal.

ii. $T_m^{(m-k)}$ is tridiagonal.

iii. The first $k-1$ components of $b_{m+1}^{(m-k)T}$ are zero.

iv. The first column of $Q_m^{(m-k)}$ is a multiple of $(A - \nu_1 I)\cdots(A - \nu_{m-k} I)q_1$.

Corollary 7.1.1 *Let* $\nu_1, \ldots, \nu_m$ *be eigenvalues of* T_m*. If the implicitly restarted QR step is performed with shifts* $\nu_1, \ldots, \nu_{m-k}$*, then the matrix* $T_m^{(m-k)}$ *has the form*

$$T_m^{(m-k)} = \begin{bmatrix} T_{kk}^{(m-k)} & T_{k,m-k}^{(m-k)} \\ 0 & T_{k+1,k+1}^{(m-k)} \end{bmatrix},$$

where $T_{k+1,k+1}^{(m-k)}$ *is an upper triangular matrix with Ritz value* $\nu_1, \ldots, \nu_{m-k}$ *on its diagonal.*

Therefore, the first k columns of the decomposition can be written in the form

$$AQ_k^{(m-k)} = Q_k^{(m-k)} T_{kk}^{(m-k)} + t_{k+1,k} q_{k+1}^{(m-k)} e_k^T + \beta_k u_{mk} q_{m+1} e_k^T,$$

where $Q_k^{(m-k)}$ consists of the first k columns of $Q_m^{(m-k)}$, $T_{kk}^{(m-k)}$ is the leading principal submatrix of order k of $T_m^{(m-k)}$, and u_{km} is from the matrix $U = U_1 \cdots U_{m-k}$. Hence if we set

$$\begin{aligned} \tilde{Q}_k &= Q_k^{(m-k)}, \\ \tilde{T}_k &= T_{kk}^{(m-k)}, \\ \tilde{\beta}_k &= \|t_{k+1,k} q_{k+1}^{(m-k)} + \beta_k u_{mk} q_{m+1}\|_2, \\ \tilde{q}_{k+1} &= \tilde{\beta}_k^{-1} (t_{k+1,k} q_{k+1}^{(m-k)} + \beta_k u_{mk} q_{m+1}), \end{aligned}$$

then

$$A\tilde{Q}_k = \tilde{Q}_k \tilde{T}_k + \tilde{\beta}_k \tilde{q}_{k+1} e_k^T$$

is a Lanczos decomposition whose starting vector is proportional to $(A - \nu_1 I) \cdots (A - \nu_{m-k} I) q_1$.

- Avoid any matrix-vector multiplications in forming the new starting vector.
- Get its Lanczos decomposition of order k for free.
- For large n the major cost will be in computing QU.

7.2 Approximation from a subspace

Assume that A is symmetric and $\{(\alpha_i, z_i)\}_{i=1}^n$ be eigenpairs of A with $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_n$. Define

$$\rho(x) = \rho(x, A) = \frac{x^T A x}{x^T x}.$$

Algorithm 7.2.1 (Rayleigh-Ritz-Quotient procedure)

Give a subspace $S^{(m)} = span\{Q\}$ with $Q^T Q = I_m$;
Set $H := \rho(Q) = Q^T A Q$;
Compute the p $(\le m)$ eigenpairs of H, which are of interest, say $Hg_i = \theta_i g_i$ for $i = 1, \ldots, p$;
Compute Ritz vectors $y_i = Qg_i$, for $i = 1, \ldots, p$;
Check $\|Ay_i - \theta_i y_i\|_2 \le Tol$, for $i = 1, \ldots, p$.

By the minimax characterization of eigenvalues, we have

$$\alpha_j = \lambda_j(A) = \min_{F^j \subseteq \mathbb{R}^n} \max_{f \in F^j} \rho(f, A).$$

Define

$$\beta_j = \min_{G^j \subseteq S^m} \max_{g \in G^j} \rho(g, A), \quad \text{for } j \leq m.$$

Since $G^j \subseteq S^m$ and $S^{(m)} = \text{span}\{Q\}$, it implies that $G^j = Q\widetilde{G}^j$ for some $\widetilde{G}^j \subseteq \mathbb{R}^m$. Therefore,

$$\beta_j = \min_{\widetilde{G}^j \subseteq \mathbb{R}^m} \max_{s \in \widetilde{G}^j} \rho(s, H) = \lambda_j(H) \equiv \theta_j,$$

for $j = 1, \ldots, m$.

For any m by m matrix B, there is associated a residual matrix $R(B) \equiv AQ - QB$.

Theorem 7.2.1 *For given orthonormal n by m matrix Q,*

$$\|R(H)\| \leq \|R(B)\|$$

for all m by m matrix B.

Proof: Since

$$\begin{aligned} R(B)^* R(B) &= Q^* A^2 Q - B^*(Q^* A Q) - (Q^* A Q)B + B^* B \\ &= Q^* A^2 Q - H^2 + (H - B)^*(H - B) \\ &= R(H)^* R(H) + (H - B)^*(H - B) \end{aligned}$$

and $(H - B)^*(H - B)$ is positive semidefinite, it implies that $\|R(B)\|^2 \geq \|R(H)\|^2$. ■

Since

$$Hg_i = \theta_i g_i, \quad \text{for } i = 1, \ldots, m,$$

we have that

$$Q^T A Q g_i = \theta_i g_i,$$

which implies that

$$QQ^T A(Qg_i) = \theta_i(Qg_i).$$

Let $y_i = Qg_i$. Then $QQ^T y_i = Q(Q^T Q)g_i = y_i$. Take $P_Q = QQ^T$ which is a projection on $\text{span}\{Q\}$. Then

$$(QQ^T)Ay_i = \theta_i(QQ^T)y_i,$$

which implies that

$$P_Q(Ay_i - \theta_i y_i) = 0,$$

i.e., $r_i = Ay_i - \theta_i y_i \perp S^m = \text{span}\{Q\}$.

Theorem 7.2.1 *Let θ_j for $j = 1, \ldots, m$ be eigenvalues of $H = Q^T AQ$. Then there is $\alpha_{j'} \in \sigma(A)$ such that*

$$|\theta_j - \alpha_{j'}| \leq \|R\|_2 = \|AQ - QH\|_2, \quad \textit{for } j = 1, \ldots, m.$$

Furthermore,

$$\sum_{j=1}^{m} (\theta_j - \alpha_{j'})^2 \leq 2\|R\|_F^2.$$

Proof: See the detail in Chapter 11 of "The symmetric eigenvalue problem , Parlett(1981)". ■

Theorem 7.2.2 *Let y be a unit vector $\theta = \rho(y)$, α be an eigenvalue of A closest to θ and z be its normalized eigenvector. Let $r = \min_{\lambda_i \neq \alpha} |\lambda_i(A) - \theta|$ and $\psi = \angle(y, z)$. Then*

$$|\theta - \alpha| \leq \|r(y)\|^2/r, \quad |\sin\psi| \leq \|r(y)\|/r,$$

where $r(y) = Ay - \theta y$.

Proof: Decompose $y = z\cos\psi + w\sin\psi$ with $z^T w = 0$ and $\|w\|_2 = 1$. Hence

$$r(y) = z(\alpha - \theta)\cos\psi + (A - \theta)w\sin\psi.$$

Since $Az = \alpha z$ and $z^T w = 0$, we have $z^T(A - \theta)w = 0$ and so

$$\|r(y)\|_2^2 = (\alpha - \theta)^2\cos^2\psi + \|(A - \theta)w\|_2^2\sin^2\psi \geq \|(A - \theta)w\|_2^2\sin^2\psi. \tag{7.2.25}$$

Let $w = \sum_{\alpha_i \neq \alpha} \xi_i z_i$. Then

$$\|(A - \theta)w\|_2^2 = |w^T(A - \theta)(A - \theta)w| = \sum_{\alpha_i \neq \alpha} (\alpha_i - \theta)^2\xi_i^2 \geq r_2(\sum_{\alpha_i \neq \alpha} \xi_i^2) = r^2.$$

Therefore,

$$|\sin\psi| \leq \frac{\|r(y)\|_2}{r}.$$

Since $r(y) \perp y$, we have

$$0 = y^T r(y) = (\alpha - \theta)\cos^2\psi + w^T(A - \theta)w\sin^2\psi,$$

which implies that

$$k \equiv \frac{\cos^2\psi}{\sin^2\psi} = \frac{w^T(A - \theta)w}{\theta - \alpha}.$$

From above equation, we get

$$sin^2\psi = \frac{1}{k+1} = \frac{\alpha - \theta}{w^T(A - \alpha)w}, \quad \cos^2\psi = \frac{k}{k+1} = \frac{w^T(A - \theta)w}{w^T(A - \alpha)w}. \tag{7.2.26}$$

Substituting (7.2.26) into (7.2.25), $\|r(y)\|_2^2$ can be rewritten as

$$\|r(y)\|_2^2 = (\theta-\alpha)w^T(A-\alpha)(A-\theta)w/w^T(A-\alpha)w. \tag{7.2.27}$$

By assumption there are no eigenvalues of A separating α and θ. Thus $(A-\alpha I)(A-\theta I)$ is positive definite and so

$$\begin{aligned} w^T(A-\alpha)(A-\theta)w &= \sum_{\alpha_i\neq\alpha} |\alpha_i-\alpha||\alpha_i-\theta|\xi_i^2 \\ &\geq r\sum_{\alpha_i\neq\alpha}|\alpha_i-\alpha|\xi_i^2 \\ &\geq r|\sum_{\alpha_i\neq\alpha}(\alpha_i-\alpha)\xi_i^2| = r|w^T(A-\alpha)w|. \end{aligned} \tag{7.2.28}$$

Substituting (7.2.28) into (7.2.27), the theorem's first inequality appears. ■

100 years old and still alive : Eigenvalue problems
Hank / G. Gloub / Van der Vorst / 2000

7.2.1 A priori bounds for interior Ritz approximations

Given subspace $S^m = \text{span}\{Q\}$, let $\{(\theta_i, y_i)\}_{i=1}^m$ be Ritz pairs of $H = Q^TAQ$ and $Az_i = \alpha_i z_i$, $i = 1,\ldots,n$.

Lemma 7.2.3 *For each $j \leq m$ for any unit $s \in S^m$ satisfying $s^Tz_i = 0$, $i = 1,\ldots,j-1$. Then*

$$\begin{aligned} \alpha_j \leq \theta_j &\leq \rho(s) + \sum_{i=1}^{j-1}(\alpha_{-1}-\theta_i)\sin^2\psi_i \\ &\leq \rho(s) + \sum_{i=1}^{j-1}(\alpha_{-1}-\alpha_i)\sin^2\psi_i, \end{aligned} \tag{7.2.29}$$

where $\psi_i = \angle(y_i, z_i)$.

Proof: Take

$$s = t + \sum_{i=1}^{j-1} r_i y_i,$$

where $t \perp y_i$ for $i = 1,\ldots,j-1$ and $\|s\|_2 = 1$. Assumption $s^Tz_i = 0$ for $i = 1,\ldots,j-1$ and

$$\begin{aligned} \|y_i - z_i\cos\psi_i\|_2^2 &= (y_i - z_i\cos\psi_i)^T(y_i - z_i\cos\psi_i) \\ &= 1 - \cos^2\psi_i - \cos^2\psi_i + \cos^2\psi_i \\ &= 1 - \cos^2\psi_i = \sin^2\psi_i \end{aligned}$$

lead to

$$|r_i| = |s^Ty_i| = |s^T(y_i - z_i\cos\psi)| \leq \|s\|_2|\sin\psi|.$$

Let (θ_i, g_i) for $i = 1, \ldots, m$ be eigenpairs of symmetric H with $g_i^T g_k = 0$ for $i \neq k$ and $y_i = Qg_i$. Then

$$0 = g_i^T (Q^T A Q) g_k = y_i^T A y_k \quad \text{for} \quad i \neq k. \tag{7.2.30}$$

Combining (7.2.30) with $t^T A y_i = 0$, we get

$$\rho(s) = t^T A t + \sum_{i=1}^{j-1} (y_i^T A y_i) r_i^2.$$

Thus

$$\begin{aligned} \rho(s) - \alpha_{-1} &= t^T (A - \alpha_{-1}) t + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1}) r_i^2 \\ &\geq \frac{t^T (A - \alpha_{-1}) t}{t^T t} + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1)} r_i^2 \\ &\geq \rho(t) - \alpha_{-1} + \sum_{i=1}^{j-1} (\theta_i - \alpha_{-1}) \sin^2 \varphi_i. \end{aligned}$$

Note that $\rho(t) \geq \theta_j = \min\{\rho(u); u \in S^{(m)}, u \perp y_i, i < j\}$. Therefore, the second inequality in (7.2.29) appears. ■

Let $\varphi_{ij} = \angle(z_i, y_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Then $\varphi_{ii} = \varphi_i$ and

$$y_j = \sum_{i=1}^{n} z_i \cos \varphi_{ij} \tag{7.2.31}$$

$$|\cos \varphi_{ij}| \leq |\sin \varphi_i| \tag{7.2.32}$$

$$\sum_{i=j+1}^{n} \cos^2 \varphi_{ij} = \sin^2 \varphi_j - \sum_{i=1}^{j-1} \cos^2 \varphi_{ij} \tag{7.2.33}$$

Proof: Since $y_j^T y_i = 0$ for $i \neq j$ and

$$|(y_i \cos \varphi_i - z_i)^T (y_i \cos \varphi_i - z_i)| = \sin^2 \varphi_i,$$

we have

$$\begin{aligned} |\cos \varphi_{ij}| &= |y_j^T z_i| = |y_j^T (y_i \cos \varphi_i - z_i)| \\ &\leq \|y_j\|_2 \|y_i \cos \varphi_i - z_i\|_2 \leq |\sin \varphi_i|. \end{aligned}$$

From (7.2.31),

$$1 = (y_j, y_j) = \sum_{i=1}^{n} \cos^2 \varphi_{ij}$$

which implies that

$$\sin^2 \varphi_j = 1 - \cos^2 \varphi_{jj} = \sum_{i=1}^{j-1} \cos^2 \varphi_{ij} + \sum_{i=j+1}^{n} \cos^2 \varphi_{ij}. \tag{7.2.34}$$

Lemma 7.2.4 *For each $j = 1, \ldots, m$,*

$$\sin\varphi_j \le [(\theta_j - \alpha_j) + \sum_{i=1}^{j-1} (\alpha_{j+1} - \alpha_i)\sin^2\varphi_i]/(\alpha_{j+1} - \alpha_j) \tag{7.2.35}$$

Proof: By (7.2.31),

$$\rho(y_j, A - \alpha_j I) = \theta_j - \alpha_j = \sum_{i=1}^{n} (\alpha_i - \alpha_j)\cos^2\varphi_{ij}.$$

It implies that

$$\begin{aligned}
& \theta_j - \alpha_j + \sum_{i=1}^{j-1} (\alpha_j - \alpha_i)\cos^2\varphi_{ij} \\
= & \sum_{i=j+1}^{n} (\alpha_i - \alpha_j)\cos^2\varphi_{ij} \\
\ge & (\alpha_{j+1} - \alpha_j)\sum_{i=j+1}^{n} \cos^2\varphi_{ij} \\
= & (\alpha_{j+1} - \alpha_j)(\sin^2\varphi_j - \sum_{i=1}^{j-1} \cos^2\varphi_{ij}). \quad \text{(from (7.2.35))}
\end{aligned}$$

Solve $\sin^2\varphi_j$ and use (9.3.10) to obtain inequality (7.2.35) ■

Explanation: By Lemmas 7.2.3 and 7.2.4, we have

$$\begin{aligned}
& j = 1: \quad \theta_1 \le \rho(s),\ s^T z_1 = 0. \quad \text{(Lemma 7.2.3)} \\
& j = 1: \quad \sin^2\varphi_1 \le \frac{\theta_1 - \alpha_1}{\alpha_2 - \alpha_1} \le \frac{\rho(s) - \alpha_1}{\alpha_2 - \alpha_1},\ s^T z_1 = 0. \quad \text{(Lemma 7.2.4)} \\
& j = 2: \quad \theta_2 \le \rho(s) + (\alpha_{-1} - \alpha_1)\sin^2\varphi_1 \le \rho(s) + (\alpha_{-1} - \alpha_1)\frac{\rho(\xi) - \alpha_1}{\alpha_2 - \alpha_1}, \\
& \qquad\quad s^T z_1 = s^T z_2 = 0,\ \xi^T z_1 = 0. \quad \text{(Lemma 7.2.3)}
\end{aligned}$$

$$\begin{aligned}
j = 2: \quad \sin^2\varphi_2 & \overset{\text{(Lemma7.2.4)}}{\le} (\theta_2 - \alpha_2) + \frac{(\alpha_3 - \alpha_1)\sin^2\varphi_1}{\alpha_3 - \alpha_2} \\
& \overset{j=1,j=2}{\le} [\rho(s) + (\alpha_{-1} - \alpha_1)(\frac{\rho(t) - \alpha_1}{\alpha_2 - \alpha_1}) - \alpha_2] + \frac{\alpha_3 - \alpha_1}{\alpha_3 - \alpha_2}(\frac{\rho(t) - \alpha_1}{\alpha_2 - \alpha_1}) \\
& \quad \vdots
\end{aligned}$$

7.3 Krylov subspace

Definition 7.3.1 *Given a nonzero vector f, $K^m(f) = [f, Af, \ldots, A^{m-1}f]$ is called Krylov matrix and $S_m = \mathcal{K}^m(f) = \langle f, Af, \ldots, A^{m-1}f\rangle$ is called Krylov subspace which are created by Lanczos if A is symmetric or Arnoldi if A is unsymmetric.*

Lemma 7.3.1 *Let $\{(\theta_i, y_i)\}_{i=1}^m$ be Ritz pairs of $K^m(f)$. If ω is a polynomial with degree $m-1$ (i.e., $\omega \in \mathcal{P}^{m-1}$), then $\omega(A)f \perp y_k$ if and only if $\omega(\theta_k) = 0,\ k = 1, \ldots, m$.*

Proof: " $\Leftarrow$ " Let

$$\omega(\xi) = (\xi - \theta_k)\pi(\xi),$$

where $\pi(\xi) \in \mathcal{P}^{m-2}$. Thus

$$\pi(A)f \in K^m(f)$$

and

$$\begin{aligned} y_k^T \omega(A) f &= y_k^T (A - \theta_k)\pi(A) f \\ &= r_k^T \pi(A) f \\ &= 0. \quad (\because r_k \perp \langle Q \rangle = \mathcal{K}^m(f)) \end{aligned}$$

" $\Rightarrow$ " exercise! ■

Define

$$\mu(\xi) \equiv \prod_{i=1}^{m} (\xi - \theta_i) \quad \text{and} \quad \pi_k(\xi) \equiv \frac{\mu(\xi)}{(\xi - \theta_k)}.$$

Corollary 7.3.2

$$y_k = \frac{\pi_k(A) f}{\|\pi_k(A) f\|}.$$

Proof: Since $\pi_k(\theta_i) = 0$ for $\theta_i \neq \theta_k$, from Lemma 7.3.1,

$$\pi_k(A) f \perp y_i,\ \forall\ i \neq k.$$

Thus, $\pi_k(A)f \ /\!/\ y_k$ and then $y_k = \frac{\pi_k(A)f}{\|\pi_k(A)f\|}$. ■

Lemma 7.3.3 *Let h be the normalized projection of f orthogonal to Z^j, $Z^j \equiv span(z_1, \ldots, z_j)$. For each $\pi \in \mathcal{P}^{m-1}$ and each $j \leq m$,*

$$\rho(\pi(A)f, A - \alpha_j I) \leq (\alpha_n - \alpha_j) \left[\frac{\sin \angle(f, Z^j)}{\cos \angle(f, Z^j)} \frac{\|\pi(A)h\|}{|\pi(\alpha_j|)} \right]^2. \tag{7.3.36}$$

Proof: Let $\psi = \angle(f, Z^j) = \cos^{-1} \|f^* Z^j\|$ and let g be the normalized projection of f onto Z^j so that

$$f = g \cos\psi + h \sin\psi.$$

Since $Z^j$is invariant under A,

$$s \equiv \pi(A)f = \pi(A)g \cos\psi + \pi(A)h \sin\psi,$$

where $\pi(A)g \in Z^j$ and $\pi(A)h \in (Z^j)^\perp$. A little calculation yields

$$\rho(s, A - \alpha_j I) = \frac{g^*(A - \alpha_j I)\pi^2(A) g \cos^2\psi + h^*(A - \alpha_j I)\pi^2(A) h \sin^2\psi}{\|\pi(A)f\|^2}. \tag{7.3.37}$$

The eigenvalues of A are labeled so that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ and

(a) $v^*(A-\alpha_j I)v \le 0$ for all $v \in Z^j$, in particular, $v = \pi(A)g$;

(b) $w^*(A-\alpha_j I)w \le (\alpha_n - \alpha_j)\|w\|^2$ for all $w \in (Z^j)^\perp$, in particular, $w = \pi(A)h$.

Used (a) and (b) to simplify (7.3.37), it becomes

$$\rho(s, A-\alpha_j I) \le (\alpha_n - \alpha_j)\left[\frac{\|\pi(A)h\|\sin\psi}{\|\pi(A)f\|}\right]^2.$$

The proof is completed by using

$$\|s\|^2 = \|\pi(A)f\|^2 = \sum_{i=1}^{n} \pi^2(\alpha_i)\cos^2\angle(f, z_i) \ge \pi^2(\alpha_j)\cos^2\angle(f, z_j).$$

■

7.3.1 The Error Bound of Kaniel and Saad

The error bounds come from choosing $\pi \in \mathcal{P}^{m-1}$ in Lemma 7.3.3 such that

(i) $|\pi(\alpha_j)|$ is large, while $\|\pi(A)h\|$ is small as possible, and

(ii) $\rho(s, A-\alpha_j I) \ge 0$ where $s = \pi(A)f$.

To (i): Note that

$$\|\pi(A)h\|^2 = \frac{\sum_{i=j+1}^{n} \pi^2(\alpha_i)\cos^2\angle(f, z_j)}{\sum_{i=j+1}^{n}\cos^2\angle(f, z_j)} \le \max_{i>j}\pi^2(\alpha_i) \le \max_{\tau\in[\alpha_{j+1},\alpha_n]}\pi^2(\tau).$$

Chebychev polynomial solves $\min_{\pi\in\mathcal{P}^{n-j}}\max_{\tau\in[\alpha_{j+1},\alpha_n]}\pi^2(\tau)$.

To (ii): The following facts are known:

(a) $0 \le \theta_j - \alpha_j$, (Cauchy interlace Theorem)

(b) $\theta_j - \alpha_j \le \rho(s, A-\alpha_j I)$, if $s \perp y_i$, for all $i < j$, (By minimax 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 $s \perp z_i$, for all $i < j$. (Lemma 7.2.3)

Theorem 7.3.4 (Saad) *Let $\theta_1 \le \cdots \le \theta_m$ be the Ritz values from $\mathcal{K}^m(f)$ (by Lanczos or Arnoldi) and let (α_i, z_i) be the eigenpairs of A. For $j = 1, \ldots, m$,*

$$0 \le \theta_j - \alpha_j \le (\alpha_n - \alpha_j)\left[\frac{\sin\angle(f, Z^j)\prod_{k=1}^{j-1}(\frac{\theta_k-\alpha_n}{\theta_k-\alpha_j})}{\cos\angle(f, Z^j)T_{m-j}(1+2r)}\right]^2$$

and

$$\tan\angle(z_j, \mathcal{K}^m) \le \frac{\sin\angle(f, Z^j)\prod_{k=1}^{j-1}(\frac{\alpha_k-\alpha_n}{\alpha_k-\alpha_j})}{\cos\angle(f, Z^j)T_{m-j}(1+2r)},$$

where $r = (\alpha_j - \alpha_{j+1})/(\alpha_{j+1} - \alpha_n)$.

Proof: Apply Lemmas 7.3.3 and 7.3.1. To ensure (b), it requires $s \perp y_i$ for $i = 1, \ldots, j-1$. By Lemma 7.3.1, we construct

$$\pi(\xi) = (\xi - \theta_1) \cdots (\xi - \theta_{j-1}) \tilde{\pi}(\xi), \quad \tilde{\pi} \in \mathcal{P}^{m-j}.$$

By Lemma 7.3.3 for this $\pi(\xi)$:

$$\begin{aligned} \frac{\|\pi(A)h\|}{|\pi(\alpha_j)|} &\leq \frac{\|(A-\theta_1)\cdots(A-\theta_{j-1})\|\|\tilde{\pi}(A)h\|}{|(\alpha_j-\theta_1)\cdots(\alpha_j-\theta_{j-1})||\tilde{\pi}(\alpha_j)|} \\ &\leq \prod_{k=1}^{j-1} \left| \frac{\alpha_n - \alpha_k}{\alpha_n - \theta_k} \right| \max_{\tau \in [\alpha_{j+1}, \alpha_j]} \frac{|\tilde{\pi}(\tau)|}{|\tilde{\pi}(\alpha_j)|} \\ &\leq \prod_{k=1}^{j-1} \left| \frac{\alpha_n - \alpha_k}{\alpha_j - \alpha_k} \right| \min_{\tilde{\pi} \in \mathcal{P}^{m-j}} \max_{j} \frac{|\tilde{\pi}(\tau)|}{|\tilde{\pi}(\alpha_j)|} \\ &= \prod_{k=1}^{j-1} \left| \frac{\alpha_n - \alpha_k}{\alpha_j - \alpha_k} \right| \frac{1}{T_{m-j}(1+2r)}. \end{aligned} \tag{7.3.38}$$

since $h \perp Z^j$. On combining (b), Lemma 7.3.3 and (7.3.38), the first of the results is obtained.
To prove the second inequality:
π is chosen to satisfy $\pi(\alpha_i) = 0$ for $i = 1, \ldots, j-1$ so that

$$s = \pi(A)f = z_j \pi(\alpha_j) \cos \angle(f, z_j) + \pi(A)h \sin \psi.$$

Therefore,

$$\tan \angle(s, z_j) = \frac{\sin \angle(f, Z^j) \|\pi(A)h\|}{\cos \angle(f, z_j) |\pi(\alpha_j)|},$$

where $\pi(\xi) = (\xi - \alpha_1) \cdots (\xi - \alpha_{j-1}) \tilde{\pi}(\xi)$ with $\tilde{\pi}(\xi) \in \mathcal{P}^{m-j}$. The proof is completed by choosing $\tilde{\pi}$ by Chebychev polynomial as above. ■

Theorem 7.3.5 *Let $\theta_{-m} \leq \ldots \leq \theta_{-1}$ be Royleigh-Ritz values of $\mathcal{K}^m(f)$ and $Az_{-j} = \alpha_{-j} z_{-j}$ for $j = n, \ldots, 1$ with $\alpha_{-n} \leq \cdots \leq \alpha_{-1}$, then*

$$0 \leq \alpha_{-j} - \theta_{-j} \leq (\alpha_{-j} - \alpha_{-1}) \left[\frac{\sin \angle(f, Z^{-j}) \prod_{k=-j+1}^{-1} (\frac{\alpha_{-n} - \theta_{-k}}{\alpha_{-k} - \theta_{-j}})}{\cos \angle(f, z_{-j}) T_{m-j}(1+2r)} \right]^2,$$

and

$$\tan(z_{-j}, \mathcal{K}^m) \leq \frac{\sin \angle(f, Z^{-j})}{\cos \angle(f, z_{-j})} \left[\frac{\prod_{k=-j+1}^{-1} (\frac{\alpha_{-k} - \alpha_{-n}}{\alpha_{-k} - \alpha_{-j}})}{T_{m-j}(1+2r)} \right]^2,$$

where $r = (\alpha_{-j-1} - \alpha_{-j})/(\alpha_{-n} - \alpha_{-j-1})$.

Theorem 7.3.6 (Kaniel) *The Rayleigh-Ritz (θ_j, y_j) from $\mathcal{K}^m(f)$ to (α_j, z_j) satisfy*

$$\begin{aligned} 0 \le \theta_j - \alpha_j \;\; &\le \;\; (\alpha_n - \alpha_j) \left[\frac{\sin \angle(f, Z^j) \prod_{k=1}^{j-1} (\frac{\alpha_k - \alpha_n}{\alpha_k - \alpha_j})}{\cos \angle(f, z_j) T_{m-j}(1+2r)} \right]^2 \\ &\quad + \sum_{k=1}^{j-1} (\alpha_n - \alpha_k) \sin^2 \angle(y_k, z_k) \end{aligned}$$

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 $r = (\alpha_j - \alpha_{j+1})/(\alpha_{j+1} - \alpha_n)$.

7.4 Applications to linear Systems and Least Squares

7.4.1 Symmetric Positive Definite System

Recall: Let A be symmetric positive definite and $Ax^* = b$. Then x^* minimizes the functional

$$\phi(x) = \frac{1}{2} x^T A x - b^T x. \tag{7.4.1}$$

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

$$x_j = Q_j y_j, \tag{7.4.2}$$

where

$$(Q_j^T A Q_j) y_j = Q_j^T b. \tag{7.4.3}$$

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:

(i) the linear system (7.4.3) must be easily solved;

(ii) we must be able to compute x_j without having to refer to $q_1, \cdots, q_j$ explicitly as (7.4.2) suggests.

To (i): we use Lanczos algorithm algorithm 7.1.1 to generate the q_i. After j steps we obtain

$$AQ_j = Q_j T_j + r_j e_j^T, \tag{7.4.4}$$

where

$$T_j = Q_j^T A Q_j = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ 0 & & \beta_{j-1} & \alpha_j \end{bmatrix} \text{ and } T_j y_j = Q_j^T b. \tag{7.4.5}$$

With this approach, (7.4.3) becomes a symmetric positive definite tridiagonal system which may be solved by LDL^T Cholesky decomposition, i.e.,

$$T_j = L_j D_j L_j^T, \tag{7.4.6}$$

where

$$L_j = \begin{bmatrix} 1 & & & 0 \\ \mu_2 & \ddots & & \vdots \\ & \ddots & \ddots & 0 \\ 0 & & \mu_j & 1 \end{bmatrix} \quad \text{and} \quad D_j = \begin{bmatrix} d_1 & & 0 \\ & \ddots & 0 \\ 0 & & d_j \end{bmatrix}.$$

Compared the entries of (7.4.6), we get

$$\begin{array}{l} d_1 = \alpha_1, \\ \text{for } \; i = 2, \cdots, j, \\ \quad \mu_i = \beta_{i-1}/d_{i-1}, \\ \quad d_i = \alpha_i - \beta_{i-1}\mu_i. \end{array} \tag{7.4.7}$$

Note that we need only calculate

$$\begin{array}{rcl} \mu_j & = & \beta_{j-1}/d_{j-1} \\ d_j & = & \alpha_j - \beta_{j-1}\mu_j \end{array} \tag{7.4.8}$$

in order to obtain L_j and D_j from L_{j-1} and D_{j-1}.
To (ii): Trick: we define $C_j = [c_1, \cdots, c_j] \in \mathbb{R}^{n \times j}$ and $p_j \in \mathbb{R}^j$ by the equations

$$\begin{array}{rcl} C_j L_j^T & = & Q_j, \\ L_j D_j p_j & = & Q_j^T b \end{array} \tag{7.4.9}$$

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 (7.4.9) 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 (7.4.8) and (7.4.9) 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 (7.4.4) that

$$Ax_j = AQ_j y_j = Q_j T_j y_j + r_j e_j^T y_j = Q_j Q_j^T b + r_j e_j^T y_j = b + r_j e_j^T y_j.$$

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 7.4.1 Given $b \in \mathbb{R}^n$ and a symmetric positive definite $A \in R^{n\times n}$. The following algorithm computes $x \in \mathbb{R}^n$ 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
    $q_{j+1} = r_j/\beta_j,$
    $\alpha_{j+1} = q_{j+1}^T A q_{j+1},$
    $\mu_{j+1} = \beta_j/d_j,$
    $d_{j+1} = \alpha_{j+1} - \mu_{j+1}\beta_j,$
    $\rho_{j+1} = -\mu_{j+1}d_j\rho_j/d_{j+1},$
    $c_{j+1} = q_{j+1} - \mu_{j+1}c_j,$
    $x_{j+1} = x_j + \rho_{j+1}c_{j+1},$
  end if
end for
$x^* = x_n.$

This algorithm requires one matrix-vector multiplication and $5n$ flops per iteration.

7.4.2 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.

7.4.3 Connection of Algorithm 7.4.1 and CG method

Let

x_j^L : Iterative vector generated by Algorithm 7.4.1

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 $i = 1, 2, \cdots$,
(a) CG method (A variant version):

$$\begin{aligned} &x_0 = 0, r_0 = b, \\ &\text{For } k = 1, \cdots, n, \\ &\quad \text{if } r_{k-1} = 0 \text{ then set } x = x_{k-1} \text{ and quit.} \\ &\quad \text{else } \beta_k = r_{k-1}^T r_{k-1} / r_{k-2}^T r_{k-2} \quad (\beta_1 \equiv 0), \\ &\qquad p_k = r_{k-1} + \beta_k p_{k-1} \quad (p_1 \equiv r_0), \\ &\qquad \alpha_k = r_{k-1}^T r_{k-1} / p_k^T A p_k, \\ &\qquad x_k = x_{k-1} + \alpha_k p_k, \\ &\qquad r_k = r_{k-1} - \alpha_k A p_k, \\ &\quad \text{end if} \\ &\text{end for} \\ &x = x_n. \end{aligned} \tag{7.4.10}$$

Define $R_k = [r_0, \cdots, r_{k-1}] \in \mathbb{R}^{n\times k}$ 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 7.4.1, 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 (7.4.11).

$$p_i^{CG} = c_i^L \cdot \text{constant}.$$

Show that c_i^L are A-orthogonal. Since

$$C_j L_j^T = Q_j \Rightarrow C_j = Q_j L_j^{-T},$$

it implies that

$$\begin{aligned} C_j^T A C_j &= L_j^{-1} Q_j^T A Q_j L_j^{-T} = L_j^{-1} T_j L_j^{-T} \\ &= L_j^{-1} L_j D_j L_j^T L_j^{-T} = D_j. \end{aligned}$$

So $\{c_i\}_{i=1}^j$ are A-orthogonal.

(b) 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 $x_j^{CG} = x_j^L$.

7.4.4 Bidiagonalization and the SVD

Suppose $U^T AV = B$ the bidiagonalization of $A \in \mathbb{R}^{m\times n}$ and that

$$\begin{array}{rcll} U & = & [u_1, \cdots, u_m], & U^T U = I_m, \\ V & = & [v_1, \cdots, v_n], & V^T V = I_n, \end{array} \tag{7.4.11}$$

and

$$B = \left[\begin{array}{cccc} \alpha_1 & \beta_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \beta_{n-1} \\ 0 & & & \alpha_n \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right]. \tag{7.4.12}$$

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

$$\begin{array}{ll} \alpha_j = \pm\|r_j\|_2, & u_j = r_j/\alpha_j, \\ v_{j+1} = p_j/\beta_j, & \beta_j = \pm\|p_j\|_2. \end{array}$$

These equations define the Lanczos method for bidiagonaling a rectangular matrix (by Paige (1974)):

```
Given v_1 ∈ R^n with unit 2-norm.
r_1 = Av_1,  α_1 = ||r_1||_2.
For j = 1, ··· , n,
    If α_j = 0 then stop;
    else
        u_j = r_j/α_j,  p_j = A^T u_j − α_j v_j,  β_j = ||p_j||_2,
        If β_j = 0  then stop;
        else
            v_{j+1} = p_j/β_j, r_{j+1} = Av_{j+1} − β_j u_j, α_{j+1} = ||r_{j+1}||_2.
        end if
    end if
end for
```
(7.4.13)

It is essentially equivalent to applying the Lanczos tridiagonalization scheme to the symmetric matrix $C = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$. 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.

7.4.5 Least square problems

As detailed in chapter III the full-rank LS problem $\min\|Ax - b\|_2$ can be solved by the bidiagonalization (7.4.11)-(7.4.12). In particular,

$$x_{LS} = V y_{LS} = \sum_{i=1}^{n} a_i v_i,$$

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:

$$U^T A V = B = \left[\begin{array}{cccc} \alpha_1 & & & 0 \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_n \\ 0 & & & \beta_n \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right], \qquad m \geq n+1,$$

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 (7.4.13). Let $V_j = [v_1, \cdots, v_j]$, $U_j = [u_1, \cdots, u_j]$ and

$$B_j = \begin{bmatrix} \alpha_1 & & & 0 \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_j \\ 0 & & & \beta_j \end{bmatrix} \in \mathbb{R}^{(j+1)\times j}$$

and consider minimizing $\|Ax - b\|_2$ over all vectors of the form $x = V_j y$, $y \in \mathbb{R}^j$. Since

$$\|AV_j y - b\|_2 = \|U^T A V_j y - U^T b\|_2 = \|B_j y - U_{j+1}^T b\|_2 + \sum_{i=j+2}^{m} (u_i^T b)^2,$$

it follows that $x_j = V_j y_j$ is the minimizer of the LS problem over span$\{V_j\}$, where y_j minimizes the $(j+1) \times j$ LS problem $\min\|B_j y - U_{j+1}^T b\|_2$. Since B_j is lower bidiagonal, it is easy to compute Jacobi rotations $J_1, \cdots, J_j$ such that

$$J_j \cdots J_1 B_j = \begin{bmatrix} R_j \\ 0 \end{bmatrix}$$

is upper bidiagonal. Let $J_j \cdots J_1 U_{j+1}^T b = \begin{bmatrix} d_j \\ u \end{bmatrix}$, then

$$\|B_j y - U_{j+1}^T b\|_2 = \|J_j \cdots J_1 y - J_j \cdots J_1 U_{j+1}^T b\|_2 = \| \begin{bmatrix} R_j \\ 0 \end{bmatrix} y - \begin{bmatrix} d_j \\ u \end{bmatrix} \|_2.$$

So $y_j = R_j^{-1} d_j$, $x_j = V_j y_j = V_j R_j^{-1} d_j = W_j d_j$. Let

$$W_j = (W_{j-1}, w_j), w_j = (v_j - w_{j-1} r_{j-1,j})/r_{jj}$$

where $r_{j-1,j}$ and r_{jj} are elements of R_j. R_j can be computed from R_{j-1}. Similarly, $d_j = \begin{bmatrix} d_{j-1} \\ \delta_j \end{bmatrix}$, x_j can be obtained from x_{j-1}:

$$x_j = W_j d_j = (W_{j-1}, w_j) \begin{bmatrix} d_{j-1} \\ \delta_j \end{bmatrix} = W_{j-1} d_{j-1} + w_j \delta_j.$$

Thus

$$x_j = x_{j-1} + w_j \delta_j.$$

For details see Paige-Saunders (1978).

7.4.6 Error Estimation of least square problems

Continuity of A^+ of the function: $\mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ defined by $A \longmapsto A^+$.

Lemma 7.4.2 *If $\{A_i\}$ converges to A and $rank(A_i) = rank(A) = n$, then $\{A_i^+\}$ also converges to A^+.*

Proof: Since $\lim_{i \to \infty} A_i^T A_i = A^T A$ 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 7.4.1 *Let $A_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \\ 0 & 0 \end{bmatrix}$ with $\varepsilon > 0$ and $A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, then $A_\varepsilon \to A_0$ as $\varepsilon \to 0$, $rank(A_0) < 2$. But $A_\varepsilon^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\varepsilon & 0 \end{bmatrix} \not\longrightarrow A_0^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ as $\varepsilon \to 0$.*

Theorem 7.4.3 *Let $A, B \in \mathbb{R}^{m\times n}$, then holds*

$$\|A^+ - B^+\|_F \leq \sqrt{2}\|A - B\|_F \max\{\|A^+\|_2^2, \|B^+\|_2^2\}.$$

Without proof.

Remark 7.4.1 *It does not follow that $A \to B$ implies $A^+ \to B^+$. Because A^+ can diverges to ∞, see example.*

Theorem 7.4.4 *If $rank(A) = rank(B)$ then*

$$\|A^+ - B^+\|_F \leq \mu\|A^+\|_2\|B^+\|_2\|A - B\|_F,$$

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

$$\begin{aligned} &A^+AA^+ = A^+, \quad (AA^+)^* = AA^+, \\ &AA^+A = A, \quad (A^+A)^* = A^+A. \end{aligned}$$

$P_A = AA^+$ 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,

$$\rho_{LS}^2 = \|b - AA^+b\|_2^2 = \|(I - AA^+)b\|_2^2.$$

Lemma 7.4.1 (Banach Lemma) $\|B^{-1} - A^{-1}\| \leq \|A - B\|\|A^{-1}\|\|B^{-1}\|$.

Proof: From $((A+\delta A)^{-1} - A^{-1})(A+\delta A) = I - I - A^{-1}\delta A$, follows lemma immediately. ■

Theorem 7.4.5 (i) *The product $P_BP_A^\perp$ can be written in the form*

$$P_BP_A^\perp = (B^+)^*R_BE^*P_A^\perp,$$

where $P_A^\perp = I - P_A$, $B = A + E$. Thus $\|P_BP_A^\perp\| \leq \|B^+\|_2\|E\|$.

(ii) *If $rank(A) = rank(B)$, then $\|P_BP_A^\perp\| \leq \min\{\|B^+\|_2, \|A^+\|_2\}\|E\|$.*

Proof:

$$\begin{aligned} P_BP_A^\perp &= P_B^*P_A^\perp = (B^+)^*B^*P_A^\perp = (B^+)^*(A+E)^*P_A^\perp = (B^+)^*E^*P_A^\perp \\ &= (B^+)^*B^*(B^+)^*E^*P_A^\perp = (B^+)^*R_BE^*P_A^\perp \quad (\|R_B\| \leq 1, \|P_A^\perp\| \leq 1). \end{aligned}$$

Part (ii) follows from the fact that $rank(A) \leq rank(B) \Rightarrow \|P_BP_A^\perp\| \leq \|P_B^\perp P_A\|$. Exercise! (Using C-S decomposition). ■

Theorem 7.4.6 *It holds*

$$B^+ - A^+ = -\overbrace{B^+P_BER_AA^+}^{F_1} + \overbrace{B^+P_BP_A^\perp}^{F_2} - \overbrace{R_B^\perp R_AA^+}^{F_3}.$$
$$B^+ - A^+ = -B^+P_BER_AA^+ + (B^*B)^+R_BE^*P_A^\perp - R_B^\perp E^*P_A(AA^*)^+.$$

Proof:

$$\begin{aligned} & -B^+BB^+(B-A)A^+AA^+ + B^+BB^+(I-AA^+) - (I-B^+B)(A^+A)A^+ \\ = \; & -B^+(B-A)A^+ + B^+(I-AA^+) - (I-B^+B)A^+ \\ = \; & B^+ - A^+ \text{ (Substitute } P_B = BB^+,\ E = B-A,\ R_A = AA^+, \cdots .). \end{aligned}$$

■

Theorem 7.4.7 *If $B = A + E$, then*

$$\|B^+ - A^+\|_F \le \sqrt{2}\|E\|_F \max\{\|A^+\|_2^2, \|B^+\|_2^2\}.$$

Proof: Suppose $rank(B) \le rank(A)$. 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 $F_1 + F_2 = B^+(P_BEA^+P_A + P_BP_A^\perp)$, we have

$$\|F_1 + F_2\|_F^2 \le \|B^+\|_2^2(\|P_BEA^+P_A\|_F^2 + \|P_BP_A^\perp\|_F^2).$$

By Theorems 7.4.5 and 7.4.6 follows that

$$\begin{aligned} \|P_BEA^+P_A\|_F^2 + \|P_BP_A^\perp\|_F^2 &\le \|P_BEA^+\|_F^2 + \|P_B^\perp P_A\|_F^2 \\ &= \|P_BEA^+\|_F^2 + \|P_B^\perp EA^+\|_F^2 \\ &= \|EA^+\|_F^2 \le \|E\|_F^2\|A^+\|_2^2. \end{aligned}$$

Thus

$$\|F_1 + F_2\|_F \le \|A^+\|_2\|B^+\|_2\|E\|_F \quad (P_B^\perp P_A = P_B^\perp ER_AA^+ = P_B^\perp EA^+).$$

By Theorem 7.4.6 we have

$$\begin{aligned} \|F_3\|_F &\le \|A^+\|_2\|R_B^\perp R_A\|_F = \|A^+\|_2\|R_AR_B^\perp\|_F = \|A^+\|_2\|A^+ER_B^\perp\| \\ &\le \|A^+\|_2^2\|E\|_F. \end{aligned}$$

The final bound is symmetric in A and B, it also holds when $rank(B) \ge rank(A)$. ■

Theorem 7.4.8 *If $rank(A) = rank(B)$, then*

$$\|B^+ - A^+\|_F \le \sqrt{2}\|A^+\|_2\|B^+\|_2\|E\|_F. \quad \textit{(see Wedin (1973))}$$

From above we have

$$\frac{\|B^+ - A^+\|_F}{\|B^+\|_2} \le \sqrt{2}k_2(A)\frac{\|E\|_F}{\|A\|_2}.$$

This bound implies that as E approaches zero, the relative error in B^+ approaches zero, which further implies that B^+ approach A^+.

Corollary 7.4.9 $\lim_{B\to A} B^+ = A^+ \iff rank(A) = rank(B)$ *as B approaches A.*

(See Stewart 1977) ■

7.4.7 Perturbation of solutions of the least square problems

We first state two corollaries of Theorem (SVD).

Theorem 7.4.10 (SVD) *If $A \in \mathbb{R}^{m\times n}$ then there exists orthogonal matrices $U = [u_1, \cdots, u_m] \in \mathbb{R}^{m\times m}$ and $V = [v_1, \cdots, v_n] \in \mathbb{R}^{n\times n}$ such that $U^T AV = diag(\sigma_1, \cdots, \sigma_p)$ where $p = min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.*

Corollary 7.4.11 *If the SVD is given by Theorem 7.4.10 and $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, then*

(a) *$rank(A) = r$.*

(b) *$\mathcal{N}(A) = span\{v_{r+1}, \cdots, v_n\}$.*

(c) *$Range(A) = span\{u_1, \cdots, u_r\}$.*

(d) *$A = \Sigma_{i=1}^r \sigma_i u_i v_i^T = U_r \Sigma_r V_r^T$, where $U_r = [u_1, \cdots, u_r]$, $V_r = [v_1, \cdots, v_r]$ and $\Sigma_r = diag(\sigma_1, \cdots, \sigma_r)$.*

(e) *$\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$.*

(f) *$\|A\|_2 = \sigma_1$.*

Proof: exercise !

Corollary 7.4.12 *Let SVD of $A \in \mathbb{R}^{m\times n}$ is given by Theorem 7.4.10. If $k < r = rank(A)$ and $A_k = \Sigma_{i=1}^k \sigma_i u_i v_i^T$, then*

$$\min_{rank(X)=k, X\in\mathbb{R}^{m\times n}} \|A - X\|_2 = \|A - A_k\|_2 = \sigma_{k+1}. \tag{7.4.14}$$

Proof: Let $X \in \mathbb{R}^{m\times n}$ 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: $\|(A + E)y - (b + f)\|_2 = \text{min!} \Rightarrow \quad y = (A + E)^+(b + f)$.

Lemma 7.4.13 *Let $A, E \in \mathbb{R}^{m\times n}$ and $rank(A) = r$.*

(a) *If $rank(A + E) > r$ then holds $\|(A + E)^+\|_2 \geq \frac{1}{\|E\|_2}$.*

(b) *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 (a): If τ_k is the smallest nonzero singular value, then $k \geq r + 1$ because of $rank(A + E) > r$. By Corollary 7.4.6, 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 (b): 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 7.4.6 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 7.4.14 *Let $A, E \in \mathbb{R}^{m\times n}$, $b, f \in \mathbb{R}^m$ 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 7.4.15 *Let $A, E \in \mathbb{R}^{m\times n}$, $b, f \in \mathbb{R}^m$, 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*

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \frac{\|A\|_2\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\left[2\frac{\|E\|_2}{\|A\|_2} + \frac{\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\frac{\|E\|_2}{\|A\|_2}\frac{\|r\|_2}{\|x\|_2} + \frac{\|f\|_2}{\|A\|_2\|x\|_2}\right].$$

Proof: From Lemma 7.4.14 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 7.4.13(b), we obtain

$$\|y-x\|_2 \leq \frac{\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\left[2\|E\|_2\|x\|_2 + \|f\|_2 + \frac{\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\|E\|_2\|r\|_2\right]$$

and

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \frac{\|A\|_2\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\left[2\frac{\|E\|_2}{\|A\|_2} + \frac{\|f\|_2}{\|A\|_2\|x\|_2} + \frac{\|A^+\|_2}{1-\|A^+\|_2\|E\|_2}\frac{\|E\|_2}{\|A\|_2}\frac{\|r\|_2}{\|x\|_2}\right].$$

■

7.5 Unsymmetric Lanczos Method

Suppose $A \in \mathbb{R}^{n\times n}$ 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

$$X = [x_1, \cdots, x_n] \text{ and } X^{-T} = Y = [y_1, \cdots, y_n].$$

Compared 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 $\alpha_j = y_j^T Ax_j$ and

$$\begin{aligned}\beta_j x_{j+1} &= \gamma_j \equiv (A-\alpha_j)x_j - \gamma_{j-1}x_{j-1}, \\ \gamma_j y_{j+1} &= p_j \equiv (A-\alpha_j)^T y_j - \beta_{j-1}y_{j-1}.\end{aligned} \tag{7.5.1}$$

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:

Algorithm 7.5.1 (Biorthogonalization method of Lanczos)

$$
\begin{array}{l}
\textit{Given } x_1, y_1 \in \mathbb{R}^n \textit{ with } x_1^T x_1 = y_1^T y_1 = 1. \\
\textit{For } j = 1, \cdots, n-1, \\
\quad \alpha_j = y_j^T A x_j, \\
\quad r_j = (A - \alpha_j) x_j - \gamma_{j-1} x_{j-1} \quad (\gamma_0 x_0 \equiv 0), \\
\quad \beta_j = \|r_j\|_2. \\
\quad \textit{If } \beta_j > 0 \textit{ then} \\
\qquad x_{j+1} = r_j / \beta_j, \\
\qquad p_j = (A - \alpha_j)^T y_j - \beta_{j-1} y_{j-1} \quad (\beta_0 y_0 \equiv 0), \\
\qquad \gamma_j = x_{j+1}^T p_j, \\
\quad \textit{else stop;} \\
\quad \textit{If } \gamma_j \neq 0 \textit{ then } y_{j+1} = p_j / \gamma_j \textit{ else stop;} \\
\textit{end for} \\
\alpha_n = x_n^T A y_n.
\end{array}
\tag{7.5.2}
$$

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

$$
\begin{aligned}
A X_j &= X_j T_j + \gamma_j e_j^T, \\
A^T Y_j &= Y_j T_j^T + p_j e_j^T.
\end{aligned}
\tag{7.5.3}
$$

Remark 7.5.1 **(i)** $p_j^T \gamma_j = \beta_j \gamma_j x_{j+1}^T y_{j+1} = \beta_j \gamma_j$ *from* (7.5.1).

(ii) *Break of the algorithm* (7.5.2) *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* (7.5.3)*).*
- **(b)** $p_j = 0 \Rightarrow \gamma_j = 0$. *Then* Y_j *is an invariant subspace of* A^T *(by* (7.5.3)*).*
- **(c)** $p_j^T \gamma_j = 0$ *but* $\|p_j\| \|\gamma_j\| \neq 0$*, then* (7.5.2) *breaks down. We begin the algorithm* (7.5.2) *with a new starting vector.*

(iii) *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* (7.5.2) *is unstable.*

Definition 7.5.1 *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 7.5.2 *Let* $A \in \mathbb{R}^{n \times n}$. *Then*

(i) *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* $H = X^{-1} A X$ *is an upper unreducible Hessenberg matrix.*

(ii) *Let* X *be a nonsingular matrix with first column* x_1 *and if* $H = X^{-1} A X$ *is an upper Hessenberg matrix, then holds*

$$K[x_1, A, n] = X K[e_1, H, n] \equiv XR,$$

where R *is an upper triangular matrix. Furthermore, if* H *is unreducible, then* R *is nonsingular.*

(iii) *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(i): 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

$$X^{-1}AX = RCR^{-1} = H$$

is an unreducible Hessenberg matrix.
ad(ii): 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(iii): Let $y_1 = \lambda x_1$. We apply (ii) to the matrix H. It follows $K[x_1, A, n] = XR_1$. Applying (ii) 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 (ii) we have R_2 is nonsingular and $X^{-1}Y = \lambda R_1 R_2^{-1} = J$ is upper triangular. So

$$\tilde{H} = Y^{-1}AY = (Y^{-1}X)X^{-1}AX(X^{-1}Y) = J^{-1}HJ.$$

■

Theorem 7.5.3 *Let $A \in \mathbb{R}^{n\times n}$, $x, y \in \mathbb{R}^n$ with $K[x, A, n]$ and $K[y, A^T, n]$ nonsingular. Then*

(i) *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.*

(ii) *Let X, Y be nonsingular 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 $\tilde{T} = D^{-1}TD$.

(iii) *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(i):

$$X = K[x, A, n]L^{-T} \Rightarrow XL^T = K[x, A, n]. \tag{7.5.4}$$

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

$$K[y, A^T, n] = K[x, A, n]^{-T} LDL^T = X^{-T}DL^T. \tag{7.5.5}$$

Applying Theorem 7.5.2(i) to (7.5.4), we get that $X^{-1}AX$ is unreducible upper Hessenberg. Applying Theorem 7.5.2(i) to (7.5.5), we get that

$$X^T A^T X^{-T} = (X^{-1}AX)^T$$

is unreducible upper Hessenberg. So $X^{-1}AX$ is an unreducible tridiagonal matrix.

ad(ii): T and $\tilde{T}$ are unreducible upper Hessenberg, by Theorem 7.5.2(3) we have $X^{-1}Y$ upper triangular on the other hand. Since $T^T = X^T A^T X^{-T}$ and $\tilde{T}^T = Y^T A^T Y^{-T}$ are unreducible upper Hessenberg, then by Theorem 7.5.2(iii) we also have $Y^T X^{-T} = (X^{-1}Y)^T$ is upper triangular. Thus $X^{-1}Y$ is upper triangular, also lower triangular so the matrix $X^{-1}Y$ is diagonal.
ad(iii): exercise! ■