5.2 Symmetric Matrices

Remark 5.10. Goal. Arnoldi's method and the minimization of the residual in $K_k\left(\underline{r}^{(0)}, A\right)$ will be studied in the special case that A is symmetric. The most important result in this case will be that it is not necessary to store the basis of $K_k\left(\underline{r}^{(0)}, A\right)$. It suffices to store a fixed number of only few basis vectors. Thus, the memory requirements do not increase in the course of the iteration and the most important problem of using GMRES vanishes. $\square$

Remark 5.11. Arnoldi's method revisited. First, Arnoldi's method is revisited. From the general relation (5.3), it follows by the orthonormality of the columns of Q_k and Q_{k+1} that

$$Q_k^T A Q_k = Q_k^T Q_{k+1} H_k = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} H_k =: \tilde{H}_k \in \mathbb{R}^{k\times k}. \tag{5.7}$$

Thus, $\tilde{H}_k$ contains just the first k rows of H_k. Since

$$\left(Q_k^T A Q_k\right)^T = Q_k^T A^T Q_k = Q_k^T A Q_k,$$

is a symmetric matrix, $\tilde{H}_k$ is symmetric, too. As in the case of a general matrix, H_k and with that $\tilde{H}_k$ is an upper Hessenberg matrix. From its symmetry, it follows that $\tilde{H}_k$ is even a tridiagonal matrix. Hence, H_k is a tridiagonal matrix, too

$$H_k = \begin{pmatrix} \alpha_1 & \beta_2 & 0 & \cdots & 0 & 0 \\ \beta_2 & \alpha_2 & \beta_3 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & & & \\ 0 & 0 & 0 & \cdots & \alpha_{k-1} & \beta_k \\ 0 & 0 & 0 & \cdots & \beta_k & \alpha_k \\ 0 & 0 & 0 & \cdots & 0 & \beta_{k+1} \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}.$$

Arnoldi's method simplifies. Using (5.3) and the special form of H_k, one obtains the relation

$$A\underline{q}_k = \beta_k \underline{q}_{k-1} + \alpha_k \underline{q}_k + \beta_{k+1}\underline{q}_{k+1}.$$

From this relation, $\underline{q}_{k+1}$ can be computed. The corresponding algorithm is called Lanczos[7] algorithm. $\square$

Algorithm 5.12. Lanczos algorithm – modified Gram–Schmidt variant. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and $\underline{q}_1 \in \mathbb{R}^n$ with $\left\|\underline{q}_1\right\|_2 = 1$.

$\beta_1 = 0$
$\underline{q}_0 = \underline{0}$
for $j = 1 : m$
  $\underline{s} = A\underline{q}_j - \beta_j \underline{q}_{j-1}$
  $\alpha_j = \left(\underline{s}, \underline{q}_j\right)$
  $\underline{s} = \underline{s} - \alpha_j \underline{q}_j$
  $\beta_{j+1} = \|\underline{s}\|_2$
  if $\beta_{j+1} == 0$

[7] Cornelius Lanczos (1893 – 1974)

```
 9.       stop
10.     endif
11.     q_{j+1} = s/β_{j+1}
12. endfor
```

□

Remark 5.13. Short recurrence. The computation of $\underline{q}_{j+1}$ requires only $\underline{q}_{j-1}$ and $\underline{q}_j$, see lines 4 – 6, and 11. This situation is called short recurrence. □

Lemma 5.14. Non-singularity of the matrix generated by the Lanczos method. *Let A be symmetric and positive definite. Then, the matrix $\tilde{H}_k = Q_k^T A Q_k$, which is generated in the Lanczos method, is non-singular.*

Proof. One has with (5.7) for all $\underline{y} \in \mathbb{R}^k$, $\underline{y} \neq \underline{0}$,

$$\underline{y}^T \tilde{H}_k \underline{y} = \underline{y}^T Q_k^T A Q_k \underline{y} = \left(Q_k \underline{y}\right)^T A \left(Q_k \underline{y}\right) > 0,$$

since Q_k has full rank and A is positive definite. Hence, $\tilde{H}_k$ is also positive definite (and symmetric). It follows that $\tilde{H}_k$ is non-singular. ■

Remark 5.15. The minimzation of the residual revisited. In the second step, one has to find a way to minimize the residual in $K_k\left(\underline{r}^{(0)}, A\right)$ without having to store the complete basis of $K_k\left(\underline{r}^{(0)}, A\right)$. Only if this is possible, then the short recurrence of the Lanczos algorithm becomes useful.

The least squares problem which has to be solved has the form (5.5). To solve this problem, a QR decomposition of H_k is used

$$H_k = \bar{Q}\bar{R}_k,\ \bar{Q} \in \mathbb{R}^{(k+1)\times(k+1)},\ \bar{R}_k \in \mathbb{R}^{(k+1)\times k}. \tag{5.8}$$

Here, $\bar{Q}$ is a unitary matrix and $\bar{R}_k$ an upper triangular matrix.

The unitary matrix $\bar{Q}$ describes, geometrically, rotations and reflections. It can be decomposed as a product of simple rotations or reflections, so-called Givens[8] rotations or Givens reflections

$$\bar{Q} = G_1 G_2 \cdots G_{k-1} G_k \tag{5.9}$$

where, in the case of a Givens rotation,

$$G_j = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c_j & s_j & & & \\ & & & -s_j & c_j & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)},\ c_j^2 + s_j^2 = 1,$$

see lecture notes Numerical Mathematics I. The non-diagonal block is in the rows/columns j and $j+1$. For a Givens reflection, the non-diagonal block has the form

$$\begin{pmatrix} c_j & s_j \\ s_j & -c_j \end{pmatrix},\ c_j^2 + s_j^2 = 1.$$

The off-diagonal entries are in the rows j and $j+1$. One gets with (5.8) and (5.9)

[8] James Wallace Givens, Jr. (1910 – 1993).

$$\bar{R}_k = \bar{Q}^T H_k = G_k^T G_{k-1}^T \cdots G_2^T G_1^T H_k.$$

Since H_k is tridiagonal, one obtains

$$\bar{R}_k = \begin{pmatrix} r_{11} & r_{12} & r_{13} & 0 & \cdots & 0 \\ 0 & r_{22} & r_{23} & r_{24} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & & & & r_{k-2,k} \\ 0 & & & & & r_{k-1,k} \\ 0 & & & & & r_{k,k} \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}, \tag{5.10}$$

i.e., $r_{ij} = 0$ if $j > i+2$. A Givens rotation changes only the rows that are involved, i.e., here the two neighboring rows j and $j+1$. The new rows are linear combinations of the old rows, where just the value zero in the entry $(j+1, j)$ is produced. A zero value can be converted to a non-zero value only at the entry $(j, j+2)$, because the entry in $(j+1, j+2)$ is generally not zero.

Consider the only interesting case $\underline{r}^{(k)} \neq \underline{0}$, in which the matrix H_k has full rank k. Let R_k be the matrix that consists of the first k rows of $\bar{R}_k$. The matrix R_k is non-singular since H_k and $\bar{Q}$ have full rank such that $\bar{R}_k$ has rank k. Setting

$$P_k = \left(\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_k\right) := Q_k R_k^{-1} \in \mathbb{R}^{n\times k},$$

one has from $P_k R_k = Q_k$ and due to the special form of R_k, compare (5.10), the recursion

$$\begin{aligned} \underline{p}_1 &= \frac{\underline{q}_1}{r_{11}}, \\ \underline{p}_2 &= \frac{1}{r_{22}}\left(\underline{q}_2 - r_{12}\underline{p}_1\right) \qquad \left(\Longleftarrow \ r_{22}\underline{p}_2 + r_{12}\underline{p}_1 = \underline{q}_2\right), \\ &\vdots \\ \underline{p}_j &= \frac{1}{r_{jj}}\left(\underline{q}_j - \underline{p}_{j-1} r_{j-1,j} - \underline{p}_{j-2} r_{j-2,j}\right), \ j = 3, \ldots, k. \end{aligned} \tag{5.11}$$

The least squares problem (5.5) can now rewritten in the form

$$\begin{aligned} &\min_{\underline{z}\in\mathbb{R}^k} \left\| \left\|\underline{r}^{(0)}\right\|_2 \underline{e}_1 - G_1 G_2 \cdots G_k \bar{R}_k \underline{z} \right\|_2^2 \\ &= \min_{\underline{z}\in\mathbb{R}^k} \left\| \left\|\underline{r}^{(0)}\right\|_2 G_k^T \cdots G_2^T G_1^T \underline{e}_1 - \bar{R}_k \underline{z} \right\|_2^2, \end{aligned}$$

because the Euclidean norm is invariant under a multiplication with a unitary matrix. Since the last row of $\bar{R}_k$ vanishes, its Moore[9]–Penrose[10] inverse (pseudo-inverse), see Numerical Mathematics I, is given by

$$\bar{R}_k^+ = \left(R_k^{-1} \ \underline{0}\right) \in \mathbb{R}^{k\times(k+1)}$$

and the solution of the least squares problem is given by

$$\underline{z}^{(k)} = \left\|\underline{r}^{(0)}\right\|_2 \bar{R}_k^+ G_k^T \cdots G_2^T G_1^T \underline{e}_1 = \left\|\underline{r}^{(0)}\right\|_2 R_k^{-1} \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le k},$$

[9] Eliakim Hastings Moore (1862 – 1932)
[10] Roger Penrose, born 1931

where the last index symbolizes that only the first k components of the vectors are taken. Consequently, the iterate with the minimal residual has the form, see (5.6),

$$\begin{aligned}\underline{x}^{(k)} &= \underline{x}^{(0)} + Q_k \underline{z}^{(k)} = \underline{x}^{(0)} + \left\|\underline{r}^{(0)}\right\|_2 Q_k R_k^{-1} (\underbrace{G_k^T \cdots G_1^T \underline{e}_1}_{\in \mathbb{R}^{k+1}})_{1\le i\le k} \\ &= \underline{x}^{(0)} + \left\|\underline{r}^{(0)}\right\|_2 P_k \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le k}.\end{aligned}$$

Since the Givens rotation or reflection G_j^T influences only the components j and $j+1$ of the vector to which it is applied, the first $(j-1)$ of its components stay unchanged:

$$\left(G_j^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le j-1} = \left(G_{j-1}^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le j-1}.$$

It follows that

$$\begin{aligned}\underline{x}^{(k)} &= \underline{x}^{(0)} + \left\|\underline{r}^{(0)}\right\|_2 P_{k-1} \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le k-1} + \left\|\underline{r}^{(0)}\right\|_2 \underline{p}_k \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{i=k} \\ &= \underbrace{\underline{x}^{(0)} + \left\|\underline{r}^{(0)}\right\|_2 P_{k-1} \left(G_{k-1}^T \cdots G_1^T \underline{e}_1\right)_{1\le i\le k-1}}_{=\underline{x}^{(k-1)}} + \left\|\underline{r}^{(0)}\right\|_2 \underline{p}_k \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{i=k} \\ &= \underline{x}^{(k-1)} + \left\|\underline{r}^{(0)}\right\|_2 \left(G_k^T \cdots G_1^T \underline{e}_1\right)_{i=k} \underline{p}_k.\end{aligned}$$

For computing $\underline{p}_k$, one needs, see (5.11), $\underline{q}_k$, $\underline{p}_{k-1}$, and $\underline{p}_{k-2}$. The result $\underline{p}_k$ can be stored in place of $\underline{p}_{k-2}$ since $\underline{p}_{k-2}$ is not needed any longer. Hence, the minimization problem can be solved without needing the complete basis of the Krylov subspace.

Together with the short recurrence of the Lanczos algorithm, it is shown that the storage of the basis of $K_k\left(\underline{r}^{(0)}, A\right)$ is not necessary.

The resulting method that computes iterates with minimal residual for symmetric matrices A is called MINRES. MINRES requires to store six arrays: $\underline{q}_k$, $\underline{q}_{k+1}$, $\underline{s}$, $\underline{p}_k$, $\underline{p}_{k-1}$, and $\underline{x}^{(k)}$. In contrast to GMRES, the current iterate $\underline{x}^{(k)}$ is known and not only the residual of the current iterate. □

Remark 5.16. S.p.d. matrices, conjugate residual method. In practice, A is often not only symmetric but also positive definite. In this case, MINRES is seldom used. Even in the context of methods that minimize the residual, there is a more efficient method for s.p.d. matrices called conjugate residual method. □

Definition 5.17. Conjugate vectors. Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. The vectors $\underline{x}, \underline{y} \in \mathbb{R}^n$ are called A-orthogonal or (A)-conjugate if

$$\underline{x}^T A \underline{y} = \left(A\underline{x}, \underline{y}\right) = 0.$$

If there is no ambiguity, the vectors are called just conjugate. □

Remark 5.18. Comparison of conjugate residual and conjugate gradient method. The conjugate residual method needs to store only five arrays. It requires in each iteration one matrix-vector product. The memory requirements are one array more than the conjugate gradient method, see Section 6.2. In addition, one has to compute one vector update ($2n$ flops) per iteration more with the conjugate residual method in comparison with the conjugate gradient method. Since both methods need in general a similar number of iterations, the conjugate gradient method is preferred in practice. For this reason, it will be refered to the literature for more details concerning the conjugate residual method. □

Remark 5.19. S.p.d. matrices vs. other matrices. As it can be already seen in the case of Krylov subspace methods that minimize the residual, one has to distinguish the cases that A is s.p.d. or A is another matrix.

These two cases represent two worlds in the context of iterative methods for solving linear systems of equations. Methods that can be used for general matrices and that are considered to work usually well in this case, are generally not among the best methods for s.p.d. matrices. The solution of systems with s.p.d. matrices is much simpler. In engineering practice, it is a common approach to try to reduce the solution of a complicated problem to the successive solution of linear systems with s.p.d. matrices. $\square$

Chapter 6
Krylov Subspace Methods that are Based on a Projection of the Residual

Remark 6.1. Idea. The methods presented in this section determine the iterate $\underline{x}^{(k)}$ at the manifold $\underline{x}^{(0)} + K_k\left(\underline{r}^{(0)}, A\right)$ such that the corresponding residual $\underline{r}^{(k)}$ is orthogonal to $K_k\left(\underline{r}^{(0)}, A\right)$. That means, $\underline{r}^{(k)}$ is projected in the orthogonal complement $K_k\left(\underline{r}^{(0)}, A\right)^{\perp}$ of $K_k\left(\underline{r}^{(0)}, A\right)$.

Orthogonality with respect to a subspace is already known as a feature of the best approximation problem in pre Hilbert spaces. There, the error is orthogonal to the subspace, see Numerical Mathematics I. In iterative schemes for solving linear equations, the error is not known. Here, one tries to construct an efficient method by projecting the residual vector. $\Box$

6.1 General Matrices

Remark 6.2. Full orthogonalization method. Let $Q_k = \{\underline{q}_1, \ldots, \underline{q}_k\}$ be an orthonormal basis of $K_k\left(\underline{r}^{(0)}, A\right)$ computed with Arnoldi's method, Algorithm 5.3. It is $\underline{q}_1 = \underline{r}^{(0)} / \left\|\underline{r}^{(0)}\right\|_2$. Set $\beta = \left\|\underline{r}^{(0)}\right\|_2$. By the orthogonality of the columns of Q_k, it follows that

$$Q_k^T \underline{r}^{(0)} = \beta Q_k^T \underline{q}_1 = \beta \underline{e}_1. \tag{6.1}$$

Then, one finds that the iterate $\underline{x}^{(k)}$ is given by

$$\underline{x}^{(k)} = \underline{x}^{(0)} + Q_k \underline{y}_k \quad \text{with} \quad \underline{y}_k = \tilde{H}_k^{-1}\left(\beta \underline{e}_1\right), \tag{6.2}$$

where $\tilde{H}_k$ is defined in (5.7), since the orthogonal projection of $\underline{r}^{(k)}$ in $K_k\left(\underline{r}^{(0)}, A\right)^{\perp}$ is unique and the iterate (6.2) fulfills $\underline{r}^{(k)} \perp K_k\left(\underline{r}^{(0)}, A\right)$:

$$\begin{aligned} Q_k^T \underline{r}^{(k)} &= Q_k^T\left(\underline{b} - A\underline{x}^{(k)}\right) = Q_k^T\left(\underline{b} - A\underline{x}^{(0)} - \beta A Q_k \tilde{H}_k^{-1} \underline{e}_1\right) \\ &= Q_k^T \underline{r}^{(0)} - \beta \underbrace{Q_k^T A Q_k}_{=\tilde{H}_k} \tilde{H}_k^{-1} \underline{e}_1 = Q_k^T \underline{r}^{(0)} - \beta \underline{e}_1 = \underline{0}, \end{aligned}$$

where (6.2), (5.7), and (6.1) have been used.

The algorithm which is based on this approach is called full orthogonalization method (FOM). Since it is of little relevance in practice, it will not be presented here in detail. Similarly to GMRES, an early break down of the Arnoldi process is equivalent of already having computed the solution. FOM possesses the same fundamental problem as GMRES since the whole basis of the $K_k\left(\underline{r}^{(0)}, A\right)$ has to be stored. In contrast to GMRES, the iterate of FOM is undefined if $\tilde{H}_k$ is singular. This situation can happen, e.g., if A is a symmetric indefinite matrix. □

6.2 Symmetric Matrices

Remark 6.3. SYMMLQ for symmetric matrices. If A is a symmetric matrix, there is a way to perform FOM with a short recurrence, i.e., without having to store the whole basis $\{\underline{q}_1, \ldots, \underline{q}_k\}$ of $K_k\left(\underline{r}^{(0)}, A\right)$. The resulting method is called SYMMLQ. This method will not be presented here. Instead, the case that A is symmetric and positive definit will be studied in detail. Then, SYMMLQ can be simplified, leading to the famous conjugate gradient (CG) method. □

Remark 6.4. Lanczos algorithm for a s.p.d. matrix. CG will be derived from the Lanczos method, Algorithm 5.12. Starting point is the Cholesky[1] decomposition of $\tilde{H}_k$, which is by Lemma 5.14 a non-singular matrix,

$$\begin{aligned}\tilde{H}_k &= L_k D_k L_k^T \\ &= \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ l_1 & 1 & \cdots & 0 & 0 \\ & \ddots & \ddots & & \\ 0 & 0 & \cdots & l_{k-1} & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & & d_k \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & l_1 & \cdots & 0 & 0 \\ & \ddots & \ddots & & \\ 0 & 0 & \cdots & 1 & l_{k-1} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.\end{aligned} \tag{6.3}$$

Define $\hat{P}_k = Q_k L_k^{-T} = \left(\underline{\hat{p}}_1, \ldots, \underline{\hat{p}}_k\right)$. The columns of $\hat{P}_k$ are linear combinations of the columns of Q_k such that $\operatorname{span}\{\underline{\hat{p}}_1, \ldots, \underline{\hat{p}}_k\} \subset K_k\left(\underline{r}^{(0)}, A\right)$. Since L_k is a non-singular matrix and since the columns of Q_k form a basis of $K_k\left(\underline{r}^{(0)}, A\right)$, the product $\hat{P}_k$ has full rank k and consequently, the columns of $\hat{P}_k$ form a basis of $K_k\left(\underline{r}^{(0)}, A\right)$. Note that $\underline{\hat{p}}_1 = \underline{q}_1 = \underline{r}^{(0)} / \left\|\underline{r}^{(0)}\right\|_2$. Using (6.3), one finds for the iterate (6.2) of FOM that

$$\underline{x}^{(k)} = \underline{x}^{(0)} + \beta Q_k L_k^{-T} D_k^{-1} L_k^{-1} \underline{e}_1 = \underline{x}^{(0)} + \hat{P}_k \underline{y}_k \tag{6.4}$$

with $\underline{y}_k = \beta D_k^{-1} L_k^{-1} \underline{e}_1$. □

Lemma 6.5. Columns of $\hat{P}_k$ are A-conjugate. *The columns $\{\underline{\hat{p}}_1, \ldots, \underline{\hat{p}}_k\}$ are mutually A-conjugate, i.e., $\hat{P}_k^T A \hat{P}_k$ is a diagonal matrix.*

Proof. Using (5.7) and (6.3) gives

$$\hat{P}_k^T A \hat{P}_k = L_k^{-1} Q_k^T A Q_k L_k^{-T} = L_k^{-1} \tilde{H}_k L_k^{-T} = D_k.$$

■

Remark 6.6. First version of a method. With this version of the method, it is shown that a short recurrence is possible. But this version is not yet optimal with respect to the computational costs.

[1] André Louis Cholesky (1875 – 1918)