Numerical Algorithms (2020) 84:1353–1380
https://doi.org/10.1007/s11075-019-00851-2

ORIGINAL PAPER

On prescribing the convergence behavior of the conjugate gradient algorithm

Gérard Meurant[1]

Received: 11 July 2019 / Accepted: 8 November 2019 / Published online: 16 December 2019

Abstract
The conjugate gradient (CG) algorithm is the most frequently used iterative method for solving linear systems $Ax = b$ with a symmetric positive definite (SPD) matrix. In this paper we construct real symmetric positive definite matrices A of order n and real right-hand sides b for which the CG algorithm has a prescribed residual norm convergence curve. We also consider prescribing as well the A-norms of the error. We completely characterize the tridiagonal matrices constructed by the Lanczos algorithm and their inverses in terms of the CG residual norms and A-norms of the error. This also gives expressions and lower bounds for the ℓ_2 norm of the error. Finally, we study the problem of prescribing both the CG residual norms and the eigenvalues of A. We show that this is not always possible. Our constructions are illustrated by numerical examples.

Keywords Conjugate gradient algorithm · Residual norm · A-norm of the error

Mathematics Subject Classification (2010) 65F10 · 65F30

1 Introduction

The conjugate gradient (CG) algorithm of Hestenes and Stiefel [11] is the algorithm of choice for solving iteratively linear systems $Ax = b$ with a symmetric positive definite (SPD) matrix. In this paper we are interested in studying different ways of constructing real SPD matrices A of order n and real right-hand sides b for which the CG algorithm has a prescribed residual norm convergence curve as well as prescribed A-norms of the error. This means that we can construct linear systems with a fast convergence of the residual norm and a slow convergence for the A-norm of the error or vice-versa. These constructions are done using the relations of CG with the

✉ Gérard Meurant
gerard.meurant@gmail.com

[1] Paris, France

Lanczos algorithm [14] and with the Full Orthogonalization method (FOM) [22, 23]. Then, we completely characterize the inverses of the tridiagonal matrices constructed by the Lanczos algorithm in terms of the CG residual norms and A-norms of the error. We also obtain expressions and lower bounds for the ℓ_2 norm of the error. Finally, we consider the problem of prescribing both the CG residual norms and the eigenvalues of A. We will show that this is not always possible. This does not come as a surprise since CG convergence depends on the eigenvalue distribution.

The well-known CG algorithm is the following,

input A, b, x_0
$r_0 = b - Ax_0$
$p_0 = r_0$
for $k = 1, \ldots$ until convergence **do**
  $\gamma_{k-1} = \frac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$
  $x_k = x_{k-1} + \gamma_{k-1} p_{k-1}$
  $r_k = r_{k-1} - \gamma_{k-1} A p_{k-1}$
  $\delta_k = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$
  $p_k = r_k + \delta_k p_{k-1}$
end for

For a summary of CG mathematical properties and its behavior in finite precision arithmetic, see [18, 19] and the references therein. The problem of prescribing the residual norms was already considered by Hestenes and Stiefel [11] in their seminal paper in 1952. Their Theorem 18.3 states that

There is no restriction whatever on the positive constants a_i, b_i (our γ_i and δ_{i+1}) in the cg-process, that is, given two sequences of positive numbers $a_0, \ldots, a_{n-1}$ and $b_0, \ldots, b_{n-1}$, there is a symmetric positive definite matrix A and a vector r_0 such that the cg-algorithm applied to A, r_0 yield the given numbers ...

and they added,

Furthermore, the formula

$$b_i = \frac{\|r_{i+1}\|^2}{\|r_i\|^2}$$

shows that there is no restriction at all on the behavior of the length of the residual vector during the cg-process.

However, given the coefficients, they did not discuss practical ways of constructing such a linear system. In this paper we show how to construct such matrices. Moreover, we also consider prescribing as well the A-norms of the error $\varepsilon_k = x - x_k$, defined as

$$\|\varepsilon_k\|_A = (A(x - x_k), x - x_k)^{\frac{1}{2}},$$

which are minimized in the CG algorithm. This implies that there exist linear systems for which the residual norm is converging fast and the A-norm of the error is converging slowly or the opposite.

From the definition of the CG algorithm it is clear that if we prescribe the residual norms we know the CG coefficients δ_k. In our notation, Hestenes and Stiefel [11] (Theorem 6.1, page 416) proved that

$$\|\varepsilon_0\|_A^2 = \|\varepsilon_k\|_A^2 + \sum_{j=0}^{k-1} \gamma_j \|r_j\|^2, \tag{1}$$

see also [25]. It implies that

$$\|\varepsilon_{k-1}\|_A^2 - \|\varepsilon_k\|_A^2 = \gamma_{k-1}\|r_{k-1}\|^2. \tag{2}$$

If we prescribe strictly decreasing values of $\|\varepsilon_k\|_A$, $k = 0, 1, \ldots, n-1$ and $\|\varepsilon_n\|_A = 0$, we obtain the coefficients $\gamma_0, \gamma_1, \ldots$ However, this does not us give us a linear system with prescribed CG residual and error norms. It can be obtained using the relations between CG and the Lanczos algorithm. It is well known that CG can be obtained from the Lanczos algorithm using the Cholesky factorization of the symmetric tridiagonal matrix generated by the Lanczos iterations. It can be written as $T = LDL^T$, using the CG coefficients with

$$L = \begin{pmatrix} 1 & & & & \\ \sqrt{\delta_1} & 1 & & & \\ & \sqrt{\delta_2} & 1 & & \\ & & \ddots & \ddots & \\ & & & \sqrt{\delta_{n-1}} & 1 \end{pmatrix} \tag{3}$$

and D is a diagonal matrix with diagonal entries $1/\gamma_0, \ldots, 1/\gamma_{n-1}$. The matrix T is symmetric tridiagonal and positive definite. We obtain a linear system with prescribed residual norms and prescribed decreasing A-norms of the error by setting $A = VTV^T$ and $b = Ve_1$ where V is any orthonormal matrix and e_1 is the first column of the identity matrix.

It seems that we are done. However, in the following, we would like to explore how what was done in the previous years for prescribing residual norms of Krylov methods for nonsymmetric problems can be used for the symmetric case.

The problem of prescribing residual norms in Krylov iterative methods for nonsymmetric matrices has been considered for quite a while (see [1, 4, 5, 9, 10, 15] and also [24]). In Section 2 we briefly recall what is known so far for the Full Orthogonalization Method (FOM) of Saad [22, 23] which reduces to CG when the matrix is SPD. In Section 3 we rely on results in [9] to construct a class of simple linear nonsymmetric systems with a prescribed FOM residual norm convergence curve. Then, we consider the symmetric case for which we construct linear systems whose matrices depend on n parameters. We show that, if some constraints on the parameters are satisfied, these matrices are positive definite. The constructed matrices are symmetric tridiagonal matrices T, obtained from their inverses, such that CG yields prescribed relative residual norms with the right-hand side e_1. Then, as above, a general SPD linear system $Ax = b$ is obtained from $A = VTV^T$ and $b = Ve_1$ where V is any orthonormal matrix. Here the right-hand side b is of unit norm but this is not a restriction. In Section 4 we show that the free parameters can be chosen to be able to prescribe the convergence curve for the A-norm of the error. In

Section 5 we completely characterize the inverses of the tridiagonal matrices produced by the Lanczos algorithm in terms of the CG residual norms and A-norms of the error. This is of interest since, in the Lanczos algorithm, the iterates are computed as $x_k = x_0 + \|r_0\| V_k T_k^{-1} e_1$ where the columns of V_k are the Lanczos basis vectors and T_k is the principal submatrix of order k of T. In Section 6 we obtain expressions and lower bounds for the ℓ_2 norm of the error in CG. In Section 7 we show that, contrary to the nonsymmetric case, it is not always possible in the symmetric case to prescribe both the residual norms and the eigenvalues of the matrix. The results of this paper are illustrated in Section 8 with four numerical examples.

2 Summary of the results for prescribing residual norms in FOM

Even though the implementation is different, FOM is mathematically equivalent to CG when the matrix is symmetric and positive definite. FOM uses an orthonormal basis of the Krylov subspace constructed with the Arnoldi process [2] which reduces to the Lanczos algorithm in the symmetric case. Prescribing the residual norms for FOM was studied some years ago. One can construct linear systems for which the matrix has prescribed eigenvalues and such that FOM delivers prescribed residual norms and also prescribed Ritz values at all iterations (see [4] and also [1, 9, 10]). If we do not have an early stop of the algorithm and if we can perform n iterations, FOM computes matrices V and H such that $AV = VH$, V being the orthonormal matrix whose columns are the basis vectors and H being an unreduced upper Hessenberg matrix with positive subdiagonal entries. Note that, in this case, the matrix A is nonderogatory. The construction can be generalized to the case for which we have an early stop (see [5]).

The upper Hessenberg matrix H can be factorized as $H = UCU^{-1}$ where U is upper triangular with $u_{1,1} = 1$ and C is the companion matrix corresponding to the (prescribed) eigenvalues of A and H (see [6]). The matrix U is given by

$$U = \left(e_1 \;\; He_1 \;\; H^2 e_1 \;\; \cdots \;\; H^{n-1} e_1 \right).$$

It was proved in [4] that the inverses of the absolute values of the entries of the first row of U^{-1} are equal to the FOM relative residual norms $\|r_k^F\|/\|r_0\|$. The other entries of U^{-1} can be chosen to prescribe the Ritz values at every iteration.

Let us denote the first row of U^{-1} as

$$\nu^T = \left(1 \;\; \hat{\nu}^T \right) = \left(\nu_1 \;\; \nu_2 \;\; \cdots \;\; \nu_n \right)^T, \quad \nu_1 = 1,$$

that is, the absolute values of the components of ν are the inverses of the relative FOM residual norms. In our constructions below, the signs of the entries of the first row of U^{-1} can be chosen arbitrarily. So, we will in general take them to be strictly positive. When we have constructed U^{-1} and C, we can compute $H = UCU^{-1}$ and the matrix of the linear system is $A = VHV^T$ where V is any orthonormal matrix. The right-hand side is $b = Ve_1$. Then, FOM starting from $x_0 = 0$ delivers the prescribed residual norms.

However, it is not easy to construct such a matrix H which is symmetric because, in that case, the matrix U must have special properties. In the next section we consider

another way to construct H and A that can be specialized to the symmetric case more easily.

3 The solution from the work of A. Greenbaum and Z. Strakoš

We first consider the nonsymmetric case and the FOM algorithm. This will provide us with the tools we need for the symmetric case.

3.1 The nonsymmetric case

Let H be an unreduced upper Hessenberg matrix. We start by constructing a nonsingular lower triangular matrix L of order n such that FOM applied to (L^{-1}, e_1) yields the same residual norms as FOM applied to (H, e_1) with $x_0 = 0$ in both cases. Note that FOM applied to (VHV^T, Ve_1) for any orthonormal matrix V yields the same residual norms.

In [9] Greenbaum and Strakoš characterized the matrices B for which the Krylov subspaces $A\mathcal{K}_k(A, v)$ and $B\mathcal{K}_k(B, v)$ are the same where v is a given vector and

$$\mathcal{K}_k(A, v) = \text{span}\{v, \quad Av, \quad A^2v, \quad \cdots \quad A^{k-1}v\}.$$

The result is the following (Theorem 1, page 8 of [9]).

Theorem 1 *Let w_i, $i = 1, \ldots, k$ be an orthonormal basis for $A\mathcal{K}_k(A, v)$ with $k \leq n$, W be the matrix whose columns are the vectors w_i, $i = 1, \ldots, n$ and H the upper Hessenberg matrix such that $AW = WH$. Then, $A\mathcal{K}_k(A, v)$ and $B\mathcal{K}_k(B, v)$ are the same for $k = 1, \ldots, n$ if and only if*

$$B = WRHW^*,$$

where R is any nonsingular upper triangular matrix.

Let us assume that $k = n$ in Theorem 1 and denote $Y = RH$. It yields $R^{-1} = HY^{-1}$. According to the theorem, the matrix HY^{-1} must be upper triangular. It is shown in [9] Theorem 3, that if H is an unreduced upper Hessenberg matrix, HX is upper triangular if and only if

$$\mathcal{L}(X) = \mathcal{L}(H^{-1})D, \tag{4}$$

where $\mathcal{L}(F)$ denotes the lower triangular matrix whose lower triangular part is the same as for the matrix F, D is a diagonal matrix and $\mathcal{L}(H^{-1})$ has no zero column. This last condition is not relevant for our problem. Relation (4) means that the lower triangular part of $X = Y^{-1}$ is the lower triangular part of H^{-1} with any column scaling.

Let us assume that we have constructed an upper Hessenberg matrix $H = UCU^{-1}$ for which FOM applied to (H, e_1) gives the prescribed residual norms we have chosen. We are looking for a nonsingular matrix L for which FOM gives the same residual norms on (L^{-1}, e_1).

From the previous results, a way to achieve this is to construct a lower triangular matrix L whose lower triangular part is the same as that of H^{-1}, taking $D = I$ in relation (4). Let us first consider what is H^{-1}, as a function of U and C, to obtain its lower triangular part. We partition U^{-1} as

$$U^{-1} = \begin{pmatrix} 1 & \hat{v}^T \\ 0 & \hat{U}^{-1} \end{pmatrix}. \tag{5}$$

We singled out the first row of U^{-1} because, as we have seen above, it is related to the FOM residual norms. It yields,

$$U = \begin{pmatrix} 1 & -\hat{v}^T\hat{U} \\ 0 & \hat{U} \end{pmatrix}.$$

Let C be the companion matrix associated with the eigenvalues of H, whose entries are

$$C = \begin{pmatrix} 0 & 0 & \dots & 0 & -\alpha_0 \\ 1 & 0 & \dots & 0 & -\alpha_1 \\ 0 & 1 & \dots & 0 & -\alpha_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & -\alpha_{n-1} \end{pmatrix}.$$

The entries of the last column are the negatives of the coefficients of the characteristic polynomial of H, α_0 being the constant term. Let $\hat{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T$. If $\alpha_0 \neq 0$ (which is the case since H is nonsingular), C is nonsingular and its inverse is

$$C^{-1} = \begin{pmatrix} -\hat{\alpha}/\alpha_0 & I_{n-1} \\ -1/\alpha_0 & 0 \end{pmatrix},$$

where I_{n-1} is the identity matrix of order $n-1$. We have to partition C^{-1} in a different way, compatible with the block structure of U and we denote the entries of the first column as

$$\beta_1 = -\alpha_1/\alpha_0, \quad \hat{\beta}^T = \begin{pmatrix} -\alpha_2/\alpha_0 & \cdots & -\alpha_{n-1}/\alpha_0 & -1/\alpha_0 \end{pmatrix}.$$

Let E be the matrix of order $n-1$ which is zero except for the first upper diagonal whose entries are equal to 1. We can write

$$C^{-1} = \begin{pmatrix} \beta_1 & e_1^T \\ \hat{\beta} & E \end{pmatrix}. \tag{6}$$

The inverse of the upper Hessenberg matrix H is characterized as follows.

Theorem 2 *Assume $\alpha_0 \neq 0$. Using the previous notation in (5) and (6) for U^{-1} and C, the inverse of H is*

$$H^{-1} = \begin{pmatrix} \beta_1 - \hat{v}^T\hat{U}\hat{\beta} & (\beta_1 - \hat{v}^T\hat{U}\hat{\beta})\hat{v}^T + (e_1^T - \hat{v}^T\hat{U}E)\hat{U}^{-1} \\ \hat{U}\hat{\beta} & \hat{U}\hat{\beta}\hat{v}^T + \hat{U}E\hat{U}^{-1} \end{pmatrix}.$$

Proof The proof is almost straightforward. We have $H^{-1} = UC^{-1}U^{-1}$. Let us first compute UC^{-1},

$$UC^{-1} = \begin{pmatrix} 1 & -\hat{v}^T\hat{U} \\ 0 & \hat{U} \end{pmatrix} \begin{pmatrix} \beta_1 & e_1^T \\ \hat{\beta} & E \end{pmatrix} = \begin{pmatrix} \beta_1 - \hat{v}^T\hat{U}\hat{\beta} & e_1^T - \hat{v}^T\hat{U}E \\ \hat{U}\hat{\beta} & \hat{U}E \end{pmatrix}.$$

Then, we right multiply by U^{-1},

$$\begin{aligned} H^{-1} &= \begin{pmatrix} \beta_1 - \hat{v}^T\hat{U}\hat{\beta} & e_1^T - \hat{v}^T\hat{U}E \\ \hat{U}\hat{\beta} & \hat{U}E \end{pmatrix} \begin{pmatrix} 1 & \hat{v}^T \\ 0 & \hat{U}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 - \hat{v}^T\hat{U}\hat{\beta} & (\beta_1 - \hat{v}^T\hat{U}\hat{\beta})\hat{v}^T + (e_1^T - \hat{v}^T\hat{U}E)\hat{U}^{-1} \\ \hat{U}\hat{\beta} & \hat{U}\hat{\beta}\hat{v}^T + \hat{U}E\hat{U}^{-1} \end{pmatrix}. \end{aligned}$$

□

We observe that the matrix $\hat{U}E\hat{U}^{-1}$ in the bottom right block of H^{-1} is strictly upper triangular that is, with a zero diagonal. Hence, Theorem 2 gives a simple proof of a result about inverses of Hessenberg matrices that was proved by several people (see, for instance, Ikebe [12]). Namely, the lower triangular part of H^{-1} is the same as the lower triangular part of a rank-one matrix. Here, we have the lower triangular part of

$$\begin{pmatrix} \beta_1 - \hat{v}^T\hat{U}\hat{\beta} \\ \hat{U}\hat{\beta} \end{pmatrix} \begin{pmatrix} 1 & \hat{v}^T \end{pmatrix}.$$

Let us construct a lower triangular matrix $L = \mathcal{L}(H^{-1})$. To simplify the notation, let $\tilde{\beta} = \beta_1 - \hat{v}^T\hat{U}\hat{\beta}$. From Theorem 2 we have

$$L = \begin{pmatrix} \tilde{\beta} & 0 \\ \hat{U}\hat{\beta} & \tilde{U} \end{pmatrix}, \tag{7}$$

with $\tilde{U} = \mathcal{L}(\hat{U}\hat{\beta}\hat{v}^T)$. The inverse of the lower triangular matrix L is

$$L^{-1} = \begin{pmatrix} \frac{1}{\tilde{\beta}} & 0 \\ -\frac{1}{\tilde{\beta}}\tilde{U}^{-1}\hat{U}\hat{\beta} & \tilde{U}^{-1} \end{pmatrix}.$$

The matrix L^{-1} is, in fact, lower bidiagonal. To prove this, let us factorize that matrix as follows.

Theorem 3 *Let $\sigma = \hat{U}\hat{\beta}$, D_v (resp. D_σ) be the diagonal matrix whose diagonal entries are $1/v_i$, $i = 1, \ldots, n$ (resp. $1/\sigma_{i-1}$ with $\sigma_0 = \tilde{\beta}$) and B_1 be the lower bidiagonal matrix*

$$B_1 = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & -1 \quad 1 \end{pmatrix}.$$

Then, the inverse of the matrix L defined by relation (7) is $L^{-1} = D_v B_1 D_\sigma$.

Proof Using the definition of $\tilde{U}$ we have $\tilde{U}^{-1}\hat{U}\hat{\beta}\hat{v}^T e_1 = e_1$. It yields

$$-\frac{1}{\tilde{\beta}}\tilde{U}^{-1}\hat{U}\hat{\beta} = -\frac{1}{\tilde{\beta}\nu_2}e_1.$$

Therefore, only the first two components of the first column of L^{-1} are nonzero.

Let us now look for the columns of $\tilde{U}^{-1}$. Let y be the solution of $\tilde{U}y = e_1$. It satisfies

$$\tilde{U}y = \begin{pmatrix} \sigma_1\nu_2 & & & \\ \sigma_2\nu_2 & \sigma_2\nu_3 & & \\ \sigma_3\nu_2 & \sigma_3\nu_3 & \sigma_3\nu_4 & \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} y = e_1.$$

The solution is given by

$$y_1 = \frac{1}{\sigma_1\nu_2}, \quad y_2 = -\frac{1}{\sigma_1\nu_3}, \quad y_j = 0,\ j = 3, \ldots, n.$$

We can compute the other columns of $\tilde{U}^{-1}$ in a similar way and we find

$$\tilde{U}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1\nu_2} & & & \\ -\frac{1}{\sigma_1\nu_3} & \frac{1}{\sigma_2\nu_3} & & \\ & -\frac{1}{\sigma_2\nu_4} & \frac{1}{\sigma_3\nu_4} & \\ & & \ddots & \ddots \end{pmatrix}.$$

Factoring the inverses of the ν_i's on the left-hand side and the inverses of the σ_i's on the right-hand side yields the result. □

We observe that the inverse of B_1 is the lower triangular matrix whose entries of the lower part are all equal to 1 and L (which is the lower triangular part of H^{-1}, as we already know from Theorem 2) is $L = D_\sigma^{-1}B_1^{-1}D_\nu^{-1}$. Therefore,

$$L = \begin{pmatrix} \sigma_0 & & & \\ \sigma_1 & \sigma_1\nu_2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{n-1} & \sigma_{n-1}\nu_2 & \cdots & \sigma_{n-1}\nu_n \end{pmatrix}. \tag{8}$$

In this matrix, the ν_i's are linked to the FOM residual norms and the σ_i's appear as parameters that we are free to choose. We observe that the ν_i's must be different from zero for L to be nonsingular. But, $\nu_k = 0$ would correspond to H_k singular and, in this case, the kth iterate of FOM is not defined. Moreover, by choosing the σ_i's appropriately we can prescribe the eigenvalues of L^{-1}.

Let us check our construction by showing that FOM applied to (L^{-1}, e_1) yields the same residual norms as FOM applied to (H, e_1). The matrix L^{-1}, which is lower bidiagonal, is the upper Hessenberg matrix obtained using the Arnoldi process with (L^{-1}, e_1), except maybe for the signs of the subdiagonal entries but this has no influence on the residual norms. To prove our claim let us consider the "U" matrix associated to L^{-1}, denoted as

$$U_{L^{-1}} = \begin{pmatrix} e_1 & L^{-1}e_1 & L^{-2}e_1 & \cdots L^{-(n-1)}e_1 \end{pmatrix}.$$

We would like to prove that the first row of $U_{L^{-1}}^{-1}$ is ν^T, that is, $e_1^T U_{L^{-1}}^{-1} = \nu^T$. This will prove that the residual norms obtained from (L^{-1}, e_1) are what we are expecting.

Theorem 4 *We have*

$$\nu^T U_{L^{-1}} = e_1^T.$$

Proof The result is obvious for the first column of $U_{L^{-1}}$. For the second column let us consider $\nu^T L^{-1}$ using the factorization of Theorem 3. First, we have $\nu^T D_\nu = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$. Then, $\nu^T D_\nu B_1 = e_n^T$. Consequently,

$$\nu^T L^{-1} = \frac{1}{\sigma_{n-1}} e_n^T.$$

Obviously, we have $\nu^T e_1 = e_1$ and $\nu^T L^{-1} e_1 = 0$. For the other columns of $U_{L^{-1}}$ we observe that

$$\nu^T L^{-j} e_1 = \nu^T L^{-1} L^{-j+1} e_1.$$

The first column of the i-th power of L^{-1} has zero last components up to $i = n-2$. It implies that $\nu^T L^{-j} e_1 = 0$ for j up to $n-1$ which proves the result. □

To summarize, applying FOM to (L^{-1}, e_1) yields relative residual norms that are equal to the inverses of the absolute values of the components of ν. The previous derivation shows that we have constructed a class of matrices and right-hand sides with a FOM prescribed convergence curve by taking $A = VL^{-1}V^T$ and $b = Ve_1$ where L is defined by relation (8) and V is any orthonormal matrix. The parameters $\sigma_0, \ldots, \sigma_{n-1}$ can be chosen as we wish. If, in addition to the residual norms, we would like to prescribe the eigenvalues of A, we can select the parameters σ_i, $i = 0, \ldots, n-1$ appropriately or we have to rely on the more complicated construction of [4] with the matrices U and C which also allows to prescribe the Ritz values. Other possibilities are to use the constructions in [1, 10, 24]. But, we observe that the construction we have described in this section is particularly simple.

3.2 The symmetric case

Our main goal in this paper is to construct symmetric matrices giving the prescribed residual norms in CG (or FOM). Following [9], we define

$$T^{-1} = \mathcal{L}(H^{-1}) + \hat{\mathcal{L}}(H^{-1})^T,$$

where $\hat{\mathcal{L}}$ gives the strict lower triangular part of the matrix (that is, with a zero diagonal). This is a kind of "symmetrization" of what we have done above. From Theorem 2 and using the same notation as before, we obtain the symmetric matrix

$$T^{-1} = \begin{pmatrix} \tilde{\beta} & (\hat{U}\hat{\beta})^T \\ \hat{U}\hat{\beta} & \mathcal{L}(\hat{U}\hat{\beta}\hat{\nu}^T) + \hat{\mathcal{L}}(\hat{U}\hat{\beta}\hat{\nu}^T)^T \end{pmatrix} = \begin{pmatrix} \sigma_0 & \sigma_1 & \cdots & \sigma_{n-1} \\ \sigma_1 & \sigma_1\nu_2 & \cdots & \sigma_{n-1}\nu_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n-1} & \sigma_{n-1}\nu_2 & \cdots & \sigma_{n-1}\nu_n \end{pmatrix}. \quad (9)$$

By computing the Cholesky factorization of T^{-1} one can see that this matrix is non-singular if $\nu_i \neq 0, i = 1, \ldots, n$ and $\nu_i\sigma_i - \nu_{i+1}\sigma_{i-1} \neq 0, i = 1, \ldots, n-1$. From the structure of T^{-1} we know that T is a symmetric tridiagonal matrix; see, for instance [8, 16]. To check our construction we would like to prove that FOM applied to (T, e_1) gives relative residual norms equal to the inverses of the absolute values of the components of ν. This will show that CG also yields the same residual norms, even though T may not be positive definite. But we will consider this problem below.

We define

$$U_T = \left(e_1 \;\; Te_1 \;\; T^2e_1 \;\; \cdots T^{n-1}e_1 \right), \tag{10}$$

and we would like to prove that $\nu^T U_T = e_1^T$.

Let μ_i, $i = 1, \ldots, n$ and η_i, $i = 1, \ldots, n-1$ be the diagonal and subdiagonal entries of the tridiagonal matrix T. By identification in the relation $T^{-1}T = I$, they can be computed from the entries of T^{-1} as follows,

$$\mu_1 = -\frac{\nu_2}{\sigma_1 - \nu_2\sigma_0}, \quad \eta_1 = \frac{1}{\sigma_1 - \nu_2\sigma_0} = -\frac{\mu_1}{\nu_2},$$

and for $i = 2, \ldots, n-1$,

$$\begin{aligned} \chi_i &= \nu_i(\nu_i\sigma_i - \nu_{i+1}\sigma_{i-1}), \\ \mu_i &= -\frac{\nu_{i+1}}{\chi_i} - \eta_{i-1}\frac{\nu_{i-1}}{\nu_i}, \quad \eta_i = \frac{\nu_i}{\chi_i}, \end{aligned} \tag{11}$$

the last diagonal entry being equal to

$$\mu_n = \frac{1}{\nu_n\sigma_{n-1}} - \eta_{n-1}\frac{\nu_{n-1}}{\nu_n}.$$

Of course, we have to assume that $\nu_i \neq 0$ for $i = 1, \ldots, n$ and $\chi_i \neq 0$ for $i = 1, \ldots, n-1$.

Lemma 1 *For T^{-1} defined in (9) and $i = 1, \ldots, n-1$, we have $\nu^T T e_i = 0$.*

Proof For the first column we obviously have

$$\mu_1 + \nu_2\eta_1 = 0.$$

For $i > 1$, the three nonzero entries of column i of T are from top to bottom $\eta_{i-1}, \mu_i, \eta_i$. Then,

$$\nu_{i-1}\eta_{i-1} + \nu_i\mu_i + \nu_{i+1}\eta_i = \nu_{i-1}\eta_{i-1} - \frac{\nu_i\nu_{i+1}}{\chi_i} - \nu_{i-1}\eta_{i-1} + \nu_{i+1}\frac{\nu_i}{\chi_i} = 0.$$

□

Theorem 5 *For U_T defined in (10) we have*

$$\nu^T U_T = e_1^T.$$

Proof The proof is similar to what we did for Theorem 4. From Lemma 1 only the last component of $\nu^T T$ is nonzero. To prove the result we just have to remark that the $(n, 1)$ entry of the ith power of T is zero for $j \leq n-2$. □

Theorem 5 proves that FOM applied to (T, e_1) yields the prescribed residual norms. For the symmetric tridiagonal matrix T, FOM is mathematically equivalent to CG. However, as we said above, T may not be positive definite and we have to find sufficient conditions to enforce the positive definiteness.

We observe that, in our construction, the values σ_i in (9) appear as parameters. Hence, the prescribed relative residual norms are obtained whatever the values of these parameters are, provided there is no division by zero. We have a general class of symmetric tridiagonal matrices giving the prescribed residual norms.

We can choose the σ_i's to obtain a positive definite matrix. This can be seen by computing the $L\Omega^{-1}L^T$ Cholesky-like factorization of T with L lower triangular and Ω diagonal. Let ω_i, $i = 1, \ldots, n$ be the diagonal entries of L and Ω.

Proposition 1 *Assume $\nu_i \neq 0$, $i = 1, \ldots, n$. The diagonal entries of the Cholesky-like factorization of T are given by*

$$\omega_i = -\frac{\nu_{i+1}}{\nu_i} \frac{1}{\nu_i\sigma_i - \nu_{i+1}\sigma_{i-1}}, \ i = 1, \ldots, n-1, \quad \omega_n = \frac{1}{\nu_n\sigma_{n-1}}.$$

Proof The proof is by induction. We have $\omega_1 = \mu_1$ and therefore, the formula is true for $i = 1$ according to the definition of μ_1. Then,

$$\omega_i = \mu_i - \frac{\eta_{i-1}^2}{\omega_{i-1}}.$$

Putting in the values of μ_i, η_{i-1} and ω_{i-1} we obtain

$$\omega_i = -\frac{\nu_{i+1}}{\chi_i} - \frac{\nu_{i-1}}{d_{i-1}}\frac{\nu_{i-1}}{\nu_i} + \frac{\nu_{i-1}^2}{d_{i-1}^2}\frac{d_{i-1}}{\nu_i} = -\frac{\nu_{i+1}}{\chi_i}.$$

Similarly, using the definition of μ_n we find that the last entry is

$$\omega_n = \frac{1}{\nu_n\sigma_{n-1}}.$$

□

As we have said above, the matrix T is nonsingular if $\nu_i \neq 0, i = 1, \ldots, n$ and $\nu_i\sigma_i - \nu_{i+1}\sigma_{i-1} \neq 0, i = 1, \ldots, n-1$. Of course, since CG is doing implicitly a factorization of the Lanczos tridiagonal matrix, we have the relation $\omega_i = 1/\gamma_{i-1}$, $i = 1, \ldots, n$. If the ω_i's are positive, the matrix T is positive definite. This is the case if the ν_i's and the σ_i' are chosen strictly positive and such that

$$\frac{\sigma_{i-1}}{\sigma_i} > \frac{\nu_i}{\nu_{i+1}}, \ i = 1, \ldots, n-1. \tag{12}$$

We will see in the next section how to choose the σ_i's to prescribe decreasing values of the A-norm of the error.

The matrix T^{-1} we have constructed, with a prescribed CG residual norm convergence curve, has as its last column a vector proportional to the vector ν. When

we apply CG to a symmetric positive definite linear system $Ax = b$, the tridiagonal matrix implicitly generated can be written as $T = UCU^{-1}$. Let us show that, conversely, the last column of T^{-1} is proportional to v.

Proposition 2 *If we denote $T = UCU^{-1}$ which is a symmetric nonsingular tridiagonal matrix, the last column of T^{-1} is proportional to the transpose of the first row of U^{-1}.*

Proof Since $T = UCU^{-1}$ is symmetric as well as T^{-1},

$$T^{-1} = T^{-T} = U^{-T}C^{-T}U^{T}.$$

The last column of U^T is $u_{n,n}e_n$. Using the notation of the present paper, the last column of C^{-T} is $-(1/\alpha_0)e_1$. Finally, we obtain

$$T^{-1}e_n = -\frac{u_{n,n}}{\alpha_0}U^{-T}e_1,$$

which proves the result. □

Theorem 6 *Let v_i, $i = 1, \ldots, n$ be given strictly positive values with $v_1 = 1$, σ_i, $i = 0, \ldots, n-1$ be strictly positive parameters satisfying condition (12) and T^{-1} be the matrix defined by (9). Define $A = VTV^T$ and $b = Ve_1$ where V is any orthonormal matrix. Then, the norms of the CG residual vectors when solving $Ax = b$ starting from $x_0 = 0$ are such that*

$$\|r_k\| = \frac{1}{v_{k+1}}, \ k = 0, 1, \ldots, n-1.$$

Note that $r_0 = b$ and $\|r_0\| = 1$. Theorem 6 shows that any residual norm convergence curve is possible for CG. In the previous construction we have $x_0 = 0$ and $\|b\| = 1$, but this is not a restriction. In the general case we would have to solve $Tx = \|r_0\|e_1$ and what is prescribed is the relative residual norm $\|r_k\|/\|r_0\|$.

4 Prescribing the *A*-norms of the error

In this section we characterize the inverses of the principal matrices T_k of T and we describe how to choose the parameters σ_i, $i = 0, \ldots, n-1$ in (9) to prescribe the A-norms of the error (or T-norms for $Tx = \|r_0\| e_1$) in addition to the prescribed residual norms. Let $\varepsilon_k = x - x_k$ be the error vector. We have the relation

$$\|\varepsilon_k\|_A^2 = \|r_0\|^2[(T^{-1})_{1,1} - (T_k^{-1})_{1,1}], \tag{13}$$

where T_k is of order k (see [17] or [8, Theorem 12.1]). To compute the difference in relation (13) we would like to relate the entries of T_k^{-1}, $k < n$ to those of T^{-1}.

We use a result proved in [13]. Theorem 2.1 of [13] says that if $B = A^{-1}$ with $a_{p,q}$ and $b_{q,p}$ different from zero and M is the submatrix of A obtained by removing row p and column q, then M is nonsingular and the entries of $C = M^{-1}$ are

$$c_{i,j} = b_{i,j} - \frac{b_{i,p} b_{q,j}}{b_{q,p}}. \tag{14}$$

In this formula, to simplify the notation, the indices (i, j) are the same as those for A and B that is, $i = 1, \ldots, q-1, q+1, \ldots, n$ and $j = 1, \ldots, p-1, p+1, \ldots, n$. But, this is not important for us since we will use this result for removing the last row and the last column. Let us first consider the principal submatrix T_{n-1} of order $n-1$.

Lemma 2 *Let T be the tridiagonal matrix whose inverse is defined by (9). The inverse of T_{n-1} is*

$$T_{n-1}^{-1} = \begin{pmatrix} \sigma_0^{(n-1)} & \sigma_1^{(n-1)} & \cdots & \sigma_{n-2}^{(n-1)} \\ \sigma_1^{(n-1)} & \sigma_1^{(n-1)}\nu_2 & \cdots & \sigma_{n-2}^{(n-1)}\nu_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n-2}^{(n-1)} & \sigma_{n-2}^{(n-1)}\nu_2 & \cdots & \sigma_{n-2}^{(n-1)}\nu_{n-1} \end{pmatrix}.$$

with

$$\sigma_i^{(n-1)} = \sigma_i - \sigma_{n-1}\frac{\nu_{i+1}}{\nu_n}, \ i = 0, 1, \ldots, n-2. \tag{15}$$

Proof We just have to consider the first and last columns of the inverse. We remove the last row and the last column. Hence, in relation (14) we have $p = q = n$. The entries of the first column are for $i = 1, \ldots, n-1$,

$$(T_{n-1}^{-1})_{i,1} = \sigma_{i-1} - \frac{(\sigma_{n-1}\nu_i)\sigma_{n-1}}{\sigma_{n-1}\nu_n} = \sigma_{i-1} - \sigma_{n-1}\frac{\nu_i}{\nu_n}.$$

The entries of the second column are for $i = 2, \ldots, n-1$,

$$(T_{n-1}^{-1})_{i,2} = \sigma_{i-1}\nu_2 - \frac{(\sigma_{n-1}\nu_i)(\sigma_{n-1}\nu_2)}{\sigma_{n-1}\nu_n} = \left(\sigma_{i-1} - \sigma_{n-1}\frac{\nu_i}{\nu_n}\right)\nu_2.$$

More generally, the entries of column j are for $i = j, \ldots, n-1$,

$$(T_{n-1}^{-1})_{i,j} = \sigma_{i-1}\nu_j - \frac{(\sigma_{n-1}\nu_i)(\sigma_{n-1}\nu_j)}{\sigma_{n-1}\nu_n} = \left(\sigma_{i-1} - \sigma_{n-1}\frac{\nu_i}{\nu_n}\right)\nu_j.$$

Hence, the entries of the last row of T_{n-1}^{-1}, which are also those of the last column are

$$\left(\sigma_{n-1}^{(n-1)} \ \sigma_{n-1}^{(n-1)}\nu_2 \ \cdots \ \sigma_{n-1}^{(n-1)}\nu_{n-1}\right),$$

using definition (15). □

To obtain the inverse of T_k, with $k = 1, \ldots, n-2$ we apply Lemma 2 recursively.

Theorem 7 *Let T be the tridiagonal matrix whose inverse is defined by (9). The inverse of T_k for $k = 1, \ldots, n-1$ is*

$$T_k^{-1} = \begin{pmatrix} \sigma_0^{(k)} & \sigma_1^{(k)} & \cdots & \sigma_{k-1}^{(k)} \\ \sigma_1^{(k)} & \sigma_1^{(k)}\nu_2 & \cdots & \sigma_{k-1}^{(k)}\nu_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k-1}^{(k)} & \sigma_{k-1}^{(k)}\nu_2 & \cdots & \sigma_{k-1}^{(k)}\nu_k \end{pmatrix}.$$

with

$$\sigma_i^{(k)} = \sigma_i^{(k+1)} - \sigma_k^{(k+1)} \frac{\nu_{i+1}}{\nu_{k+1}}, \ i = 0, 1, \ldots, k-1. \tag{16}$$

Moreover,

$$\sigma_i^{(k)} = \sigma_i - \sigma_k \frac{\nu_{i+1}}{\nu_{k+1}}, \ i = 0, 1, \ldots, k-1. \tag{17}$$

Proof Formula (16) is obtained by applying Lemma 2 to T_{k+1}. To prove relation (17), we proceed by induction from $k = n-1$ to $k = 1$. From Lemma 2, the result holds for $k = n-1$. Let us assume that it holds for $k+1$. Then, for $i = 0, \ldots, k-1$,

$$\begin{aligned} \sigma_i^{(k)} &= \sigma_i^{(k+1)} - \sigma_k^{(k+1)} \frac{\nu_{i+1}}{\nu_{k+1}}, \\ &= \sigma_i - \sigma_{k+1} \frac{\nu_{i+1}}{\nu_{k+2}} - \left(\sigma_k - \sigma_{k+1} \frac{\nu_{k+1}}{\nu_{k+2}} \right) \frac{\nu_{i+1}}{\nu_{k+1}}, \\ &= \sigma_i - \sigma_k \frac{\nu_{i+1}}{\nu_{k+1}}. \end{aligned}$$

□

Corollary 1 *Using the notation of Theorem 7, we have*

$$\sigma_i^{(k)} = \sigma_i^{(j)} - \sigma_k^{(j)} \frac{\nu_{i+1}}{\nu_{k+1}}, \ i = 0, 1, \ldots, k-1, \ n \geq j > k. \tag{18}$$

Proof The proof is similar to the one for relation (17). □

We observe that the results of Theorem 7 and Corollary 1 are valid for the inverse of any nonsingular symmetric tridiagonal matrix since the inverse can always be expressed as (9) (see [3]). Now, we apply these results to the computation of the A-norm of the error.

Theorem 8 *Using the notation of the previous section, the square of the A-norm of the error is given by*

$$\|\varepsilon_k\|_A^2 = \|r_0\|^2 \frac{\sigma_k}{\nu_{k+1}} = |\sigma_k| \, \|r_0\| \, \|r_k\|, \ k = 0, \ldots, n-1. \tag{19}$$

Proof From Theorem 7 we have

$$(T^{-1})_{1,1} - (T_k^{-1})_{1,1} = \sigma_k \frac{1}{\nu_{k+1}}.$$

Using relation (13) and the definition of ν_{k+1}, the result follows. □

Consequently, we can prescribe the values of the A-norms of the error by computing the σ_i's using relation (19). Of course, the prescribed values for $\|\varepsilon_k\|_A$ must be decreasing. This corresponds to the condition (12) being satisfied and to the matrix T being positive definite.

Relation (19) can be related to the Hestenes and Stiefel relation (1),

$$\|\varepsilon_k\|_A^2 = \sum_{j=k}^{n-1} \gamma_j \|r_j\|^2, \quad \gamma_j = \frac{\|r_j\|^2}{(Ap_j, p_j)}.$$

The coefficient γ_j is one of the two coefficients computed in CG. By dividing every term of the sum by $\|r_0\| \, \|r_k\|$ and by identification, we obtain

$$|\sigma_k| = \sum_{j=k}^{n-1} \gamma_j \frac{\|r_j\|^2}{\|r_0\| \, \|r_k\|}. \tag{20}$$

In particular, this yields $|\sigma_{n-1}| = \gamma_{n-1}\|r_{n-1}\|/\|r_0\|$ and $\|\varepsilon_{n-1}\|_A^2 = \gamma_{n-1}\|r_{n-1}\|^2$.

5 The Lanczos tridiagonal matrix and its inverse

In this section we consider the tridiagonal matrix T that is obtained from the Lanczos algorithm or implicitly from CG. We would like to characterize the entries of T and the entries of the inverse of T as functions of the residual norms and the A-norms of the error. To do this we just have to consider the first and last columns of the inverse since all the other entries can be obtained from them.

As in Section 1, let γ_j, $j = 0, \ldots, n-1$ and δ_j, $j = 1, \ldots, n-1$ be the CG coefficients. We use the LDL^T factorization of the Lanczos tridiagonal matrix we have seen in Section 1. The lower bidiagonal matrix L is given by relation (3) with diagonal entries equal to 1 and D is a diagonal matrix with diagonal entries $1/\gamma_0, \ldots, 1/\gamma_{n-1}$. Let us compute the inverse of L. For the column j of the inverse we must have $L\ell^{(j)} = e_j$. It yields

$$\begin{aligned}
&\ell_i^{(j)} = 0, \; i = 1, \ldots, j-1, \quad \ell_j^{(j)} = 1, \\
&\ell_i^{(j)} = (-1)^{i-j} \sqrt{\delta_1 \cdots \delta_{i-1}}, \; i = j+1, \ldots, n,
\end{aligned}$$

and the inverse is

$$L^{-1} = \begin{pmatrix} 1 & & & & \\ -\sqrt{\delta_1} & 1 & & & \\ \sqrt{\delta_1\delta_2} & -\sqrt{\delta_2} & 1 & & \\ -\sqrt{\delta_1\delta_2\delta_3} & \sqrt{\delta_2\delta_3} & \ddots & 1 & \\ \vdots & \vdots & & \ddots & \ddots \\ (-1)^{n-1}\sqrt{\delta_1\cdots\delta_{n-1}} & (-1)^{n-2}\sqrt{\delta_2\cdots\delta_{n-1}} & \cdots & -\sqrt{\delta_{n-1}} & 1 \end{pmatrix}$$

Then, we can compute the product $L^{-T}D^{-1}L^{-1}$. We are particularly interested in the first column $T_{:,1}^{-1}$ which is

$$L^{-T}\begin{pmatrix} \gamma_0 \\ -\gamma_1\sqrt{\delta_1} \\ \gamma_2\sqrt{\delta_1\delta_2} \\ -\gamma_3\sqrt{\delta_1\delta_2\delta_3} \\ \vdots \\ (-1)^{n-1}\gamma_{n-1}\sqrt{\delta_1\cdots\delta_{n-1}} \end{pmatrix}.$$

It yields

$$T_{:,1}^{-1} = \begin{pmatrix} \gamma_0 + \gamma_1\delta_1 + \gamma_2\delta_1\delta_2 + \cdots + \gamma_{n-1}\delta_1\cdots\delta_{n-1} \\ -[\gamma_1\sqrt{\delta_1} + \gamma_2\sqrt{\delta_1}\delta_2 + \gamma_3\sqrt{\delta_1}\delta_2\delta_3 + \cdots + \gamma_{n-1}\sqrt{\delta_1}\delta_2\cdots\delta_{n-1}] \\ \vdots \\ (-1)^{n-2}[\gamma_{n-2}\sqrt{\delta_1\cdots\delta_{n-2}} + \gamma_{n-1}\sqrt{\delta_1\cdots\delta_{n-2}}\delta_{n-1}] \\ (-1)^{n-1}\gamma_{n-1}\sqrt{\delta_1\cdots\delta_{n-1}} \end{pmatrix}.$$

This can be factored as $T_{:,1}^{-1} = \Delta t$ where Δ is a diagonal matrix with $\Delta_{1,1} = 1$, $\Delta_{i,i} = (-1)^{i-1}\sqrt{\delta_1\cdots\delta_{i-1}}$ for $i = 2, \ldots, n$ and the vector t is

$$t = \begin{pmatrix} \gamma_0 + \gamma_1\delta_1 + \gamma_2\delta_1\delta_2 + \cdots + \gamma_{n-1}\delta_1\cdots\delta_{n-1} \\ \gamma_1 + \gamma_2\delta_2 + \gamma_3\delta_2\delta_3 + \cdots + \gamma_{n-1}\delta_2\cdots\delta_{n-1} \\ \vdots \\ \gamma_{n-2} + \gamma_{n-1}\delta_{n-1} \\ \gamma_{n-1} \end{pmatrix}.$$

In CG we have $\delta_k = \|r_k\|^2/\|r_{k-1}\|^2$. Hence, for $i < j$, $\delta_i\cdots\delta_j = \|r_j\|^2/\|r_{i-1}\|^2$ and

$$\Delta_{i,i} = (-1)^{i-1}\frac{\|r_{i-1}\|}{\|r_0\|},\ i = 1, \ldots, n.$$

It yields

$$T_{i,1}^{-1} = (-1)^{i-1} \frac{1}{\|r_0\| \, \|r_{i-1}\|} \sum_{j=i-1}^{n-1} \gamma_j \, \|r_j\|^2, \ i = 1, \ldots, n,$$

whose absolute value was already known from relation (20). From Theorem 8 we know that $\|\varepsilon_i\|_A^2 = |\sigma_i| \, \|r_0\| \, \|r_i\|$ and $\sigma_i = T_{i+1,1}^{-1}$ for $i = 0, \ldots, n-1$. Therefore,

$$T_{i,1}^{-1} = (-1)^{i-1} \frac{\|\varepsilon_{i-1}\|_A^2}{\|r_0\| \, \|r_{i-1}\|}, \ i = 1, \ldots, n.$$

The last column of T^{-1} is

$$T_{:,n}^{-1} = \gamma_{n-1} \begin{pmatrix} (-1)^{n-1} \sqrt{\delta_1 \cdots \delta_{n-1}} \\ (-1)^{n-2} \sqrt{\delta_2 \cdots \delta_{n-1}} \\ \vdots \\ -\sqrt{\delta_{n-1}} \\ 1 \end{pmatrix} = \gamma_{n-1} \begin{pmatrix} (-1)^{n-1} \frac{\|r_{n-1}\|}{\|r_0\|} \\ (-1)^{n-2} \frac{\|r_{n-1}\|}{\|r_1\|} \\ \vdots \\ -\frac{\|r_{n-1}\|}{\|r_{n-2}\|} \\ 1 \end{pmatrix}$$

$$= \gamma_{n-1} \|r_{n-1}\| \begin{pmatrix} (-1)^{n-1} \frac{1}{\|r_0\|} \\ (-1)^{n-2} \frac{1}{\|r_1\|} \\ \vdots \\ -\frac{1}{\|r_{n-2}\|} \\ \frac{1}{\|r_{n-1}\|} \end{pmatrix} = \frac{\|\varepsilon_{n-1}\|_A^2}{\|r_{n-1}\|} \begin{pmatrix} (-1)^{n-1} \frac{1}{\|r_0\|} \\ (-1)^{n-2} \frac{1}{\|r_1\|} \\ \vdots \\ -\frac{1}{\|r_{n-2}\|} \\ \frac{1}{\|r_{n-1}\|} \end{pmatrix},$$

since $\|\varepsilon_{n-1}\|_A^2 = \gamma_{n-1} \|r_{n-1}\|^2$. From this and the relation

$$\sigma_{n-1} = (-1)^{n-1} \|\varepsilon_{n-1}\|_A^2 / (\|r_0\| \, \|r_{n-1}\|)$$

we obtain

$$\nu = \begin{pmatrix} 1 \\ -\frac{\|r_0\|}{\|r_1\|} \\ \vdots \\ (-1)^{n-1} \frac{\|r_0\|}{\|r_{n-1}\|} \end{pmatrix}.$$

Hence, the signs of the components of ν alternate. Since we have

$$\sigma_i = T_{i+1,1}^{-1} = (-1)^i \frac{\|\varepsilon_i\|_A^2}{\|r_0\| \, \|r_i\|}, \ i = 0, \ldots, n-1,$$

and

$$\nu_i = (-1)^{i-1} \frac{\|r_0\|}{\|r_{i-1}\|}, \ i = 1, \ldots, n,$$

we can recover the nonzero entries of T using the formula (11).

Theorem 9 *The nonzero entries of the symmetric tridiagonal matrix T implicitly generated by the CG algorithm are given by*

$$\mu_1 = \frac{\|r_0\|^2}{\|\varepsilon_0\|_A^2 - \|\varepsilon_1\|_A^2}, \quad \eta_1 = \frac{\|r_0\| \, \|r_1\|}{\|\varepsilon_0\|_A^2 - \|\varepsilon_1\|_A^2},$$

and for $i = 2, \ldots, n$,

$$\mu_i = \|r_{i-1}\|^2 \frac{\|\varepsilon_{i-2}\|_A^2 - \|\varepsilon_i\|_A^2}{(\|\varepsilon_{i-1}\|_A^2 - \|\varepsilon_i\|_A^2)\,(\|\varepsilon_{i-2}\|_A^2 - \|\varepsilon_{i-1}\|_A^2)},$$

$$\eta_i = \frac{\|r_i\| \, \|r_{i-1}\|}{\|\varepsilon_{i-1}\|_A^2 - \|\varepsilon_i\|_A^2}.$$

Proof After some computations we obtain

$$\nu_i \sigma_i - \nu_{i+1} \sigma_{i-1} = \frac{\|\varepsilon_{i-1}\|_A^2 - \|\varepsilon_i\|_A^2}{\|r_i\| \, \|r_{i-1}\|},$$

$$\frac{\nu_{i+1}}{\nu_i} = -\frac{\|r_{i-1}\|}{\|r_i\|}.$$

Inserting these expressions in the coefficients of the diagonal and subdiagonal entries of T proves the result. We observe that this result could have also be obtained by computing $T = LDL^T$ directly. □

Note that μ_i, $i = 2, \ldots, n$ can also be written as

$$\mu_i = \frac{\|r_{i-1}\|^2}{\|\varepsilon_{i-1}\|_A^2 - \|\varepsilon_i\|_A^2} + \frac{\|r_{i-1}\|^2}{\|\varepsilon_{i-2}\|_A^2 - \|\varepsilon_{i-1}\|_A^2}.$$

We also know that

$$\gamma_{i-1} = \frac{\|\varepsilon_{i-1}\|_A^2 - \|\varepsilon_i\|_A^2}{\|r_{i-1}\|^2}.$$

Hence, we have expressions for all the Lanczos and CG coefficients as functions of the residual norms and A-norms of the error.

The matrices T_k are the principal matrices of T but the entries of their inverses are, of course, different from the corresponding entries of the inverse of T. They can be computed from Theorem 7 and the relation (17). It yields

$$\sigma_i^{(k)} = (-1)^i \frac{1}{\|r_0\| \, \|r_i\|} (\|\varepsilon_i\|_A^2 - \|\varepsilon_k\|_A^2), \; i = 0, 1, \ldots, k-1.$$

We observe that if $\|\varepsilon_k\|_A^2$ is small, $\sigma_i^{(k)}$ must be close to σ_i, the corresponding entry in T^{-1}. The relative difference is

$$\left| \frac{\sigma_i^{(k)} - \sigma_i}{\sigma_i} \right| = \frac{\|\varepsilon_k\|_A^2}{\|\varepsilon_i\|_A^2} < 1, \; 0 \le i < k.$$

The convergence of the entries of the first column of T_k^{-1} to the corresponding ones of the first column of T^{-1} is therefore linked to the convergence of the A-norm of the error.

6 The ℓ_2 norm of the error

The results of the previous sections can be used to obtain an expression of the error $\varepsilon_k = x - x_k$. For simplicity we assume that CG terminates at iteration n giving the relation $AV = VT$.

Theorem 10 *Let us assume that A has distinct eigenvalues and that CG terminates at iteration n. The error vector is*

$$\varepsilon_k = \|r_0\| \, V \begin{pmatrix} \frac{\sigma_k}{\nu_{k+1}} \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_k \end{pmatrix} \\ \sigma_k \\ \vdots \\ \sigma_{n-1} \end{pmatrix}. \tag{21}$$

Proof The exact solution is given by $x = x_0 + Vy$ where $y = \|r_0\| T^{-1} e_1$ whence $x_k = x_0 + V_k y^{(k)}$ with $y^{(k)} = \|r_0\| T_k^{-1} e_1$ and V_k is the matrix of the first k columns of V. Therefore,

$$\varepsilon_k = \|r_0\| \, V \left[T^{-1} e_1 - \begin{pmatrix} T_k^{-1} e_1 \\ 0 \end{pmatrix} \right],$$

with a slight abuse of notation since the two vectors e_1 do not have the same dimension. It yields

$$\varepsilon_k = \|r_0\| \, V \left[\begin{pmatrix} \sigma_0 \\ \vdots \\ \vdots \\ \vdots \\ \sigma_{n-1} \end{pmatrix} - \begin{pmatrix} \sigma_0^{(k)} \\ \vdots \\ \sigma_{k-1}^{(k)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right].$$

The result follows by using relation (17) of Theorem 7. □

From Theorem 10 we obtain the ℓ_2 norm of the error.

Corollary 2

$$\|\varepsilon_k\|^2 = \|\varepsilon_k\|_A^4 \sum_{j=0}^{k-1} \frac{1}{\|r_j\|^2} + \sum_{j=k}^{n-1} \frac{\|\varepsilon_j\|_A^4}{\|r_j\|^2},$$

and we have a lower bound for the CG ℓ_2 norm of the error,

$$\|\varepsilon_k\|_A^4 \sum_{j=0}^{k-1} \frac{1}{\|r_j\|^2} \leq \|\varepsilon_k\|^2. \tag{22}$$

Proof We just have to use the fact that V is an orthonormal matrix and the relations we have for the ν_j's and σ_j's to obtain the result from Theorem 10. □

Note that we have

$$\sum_{j=0}^{k-1} \frac{1}{\|r_j\|^2} = \frac{\|p_{k-1}\|^2}{\|r_{k-1}\|^4},$$

see, for instance [18], page 53.

Of course, at iteration k we do not know $\|\varepsilon_k\|_A$ but, to obtain a computable lower bound of the ℓ_2 error norm, we can replace the A-norm of the error by a lower bound computed using Gauss quadrature (see [8]).

7 Can we prescribe the residual norms and the eigenvalues of *A*?

Let λ_i, $i = 1, \ldots, n$ be real strictly positive distinct numbers. They are the eigenvalues we would like to eventually prescribe and let C be the corresponding companion matrix. Let $T = UCU^{-1}$ with U upper triangular. This is an unreduced upper Hessenberg matrix. We would like to have T symmetric with a prescribed first row of U^{-1}. This will give a symmetric tridiagonal matrix with given eigenvalues and prescribed residual norms when solving $Tx = e_1$. Hence, we must have $UCU^{-1} = U^{-T}C^TU^T$. This can be written as

$$UCU^{-1} = U^{-T}C^TU^T \Rightarrow U^TUC = C^TU^TU. \tag{23}$$

The problem on the right-hand side of (23) is a Sylvester equation but, C and C^T have the same eigenvalues and the solution U^TU is not unique. This problem was studied by Fiedler in [7]. His notation is different from ours. In particular, the companion matrix he considered is the transpose of ours. But, stated in our notation and specialized to our needs, his result (Theorem 2.1) is the following.

Theorem 11 *Let p be the monic polynomial whose roots are λ_i, $i = 1, \ldots, n$, C the associated companion matrix and $\mathcal{V}$ the Vandermonde matrix constructed with the λ_i's (with ones on the first column). The following statements are equivalent,*

1. *$\mathcal{H}$ is a Hankel matrix compatible with p,*
2. *$\mathcal{H}C = C^T\mathcal{H}$,*
3. *$\mathcal{V}^{-T}\mathcal{H}\mathcal{V}^{-1}$ is a diagonal matrix D.*

If the Hankel matrix is denoted as

$$\mathcal{H} = \begin{pmatrix} h_0 & h_1 & \cdots & h_{n-1} \\ h_1 & h_2 & \cdots & h_n \\ \vdots & \vdots & \ddots & \vdots \\ h_{n-1} & h_n & \cdots & h_{2n-2} \end{pmatrix},$$

and the polynomial p is $p_n\lambda^n + \cdots + p_1\lambda + p_0$, $\mathcal{H}$ is compatible with p if

$$\begin{pmatrix} h_0 & h_1 & \cdots & h_{n-1} & h_n \\ h_1 & h_2 & \cdots & h_n & h_{n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{n-2} & h_{n-1} & \cdots & h_{2n-3} & h_{2n-2} \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_n \end{pmatrix} = 0 \tag{24}$$

is satisfied. In our case $p_n = 1$. The theorem proved by Fiedler is more general than Theorem 11 since it allows to have multiple roots for the polynomial and a confluent Vandermonde matrix. Theorem 11 corresponds to the solution of the equality (23) for UCU^{-1}. It tells us that U^TU is the Cholesky factorization of a Hankel matrix $\mathcal{H} = \mathcal{V}^T D\mathcal{V}$. The question is to know if there is a matrix D with a positive diagonal such that ν can be the first row of U^{-1}.

Let

$$U^{-1} = \begin{pmatrix} 1 & \hat{\nu}^T \\ 0 & \hat{U}^{-1} \end{pmatrix}.$$

Then, using relation 3 in Theorem 11, we must have

$$\begin{aligned} U^TU &= \begin{pmatrix} 1 & -\hat{\nu}^T\hat{U} \\ -\hat{U}^T\hat{\nu} & \hat{U}^T\hat{U} + \hat{U}^T\hat{\nu}\hat{\nu}^T\hat{U} \end{pmatrix} \\ &= \begin{pmatrix} d_1 & d_2 & \cdots & d_n \\ d_1\lambda_1 & d_2\lambda_2 & \cdots & d_n\lambda_n \\ \vdots & \vdots & & \vdots \\ d_1\lambda_1^{n-1} & d_2\lambda_2^{n-1} & \cdots & d_n\lambda_n^{n-1} \end{pmatrix} \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & & \vdots & \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{pmatrix}. \end{aligned}$$

Looking at the first column, we must have

$$\sum_{j=1}^{n} d_j = 1, \quad [\hat{U}^T\hat{\nu}]_i = -\sum_{j=1}^{n} d_j\lambda_j^i, \quad i = 1, \ldots, n-1.$$

The second relation can be used for the bottom right term. Given ν, in total we have $(n^2+n)/2$ unknowns that is, the nonzero entries in $\hat{U}$ and the diagonal entries of D, and the same number of equations. However, we have a polynomial system and it is not obvious that we can find a real solution with $d_j > 0,\ j = 1, \ldots, n$. In fact, this is not always possible.

To see this, let us consider the case $n = 2$. We have

$$U^TU = \begin{pmatrix} 1 & -u_{2,2}\nu_2 \\ -u_{2,2}\nu_2 & u_{2,2}^2(1+\nu_2^2) \end{pmatrix} = \mathcal{V}^T D\mathcal{V} = \begin{pmatrix} d_1 + d_2 & d_1\lambda_1 + d_2\lambda_2 \\ d_1\lambda_1 + d_2\lambda_2 & d_1\lambda_1^2 + d_2\lambda_2^2 \end{pmatrix}.$$

We have three equations and three unknowns. From the (1, 1) entry we can eliminate $d_2 = 1 - d_1$. Then, we have

$$d_1(\lambda_1 - \lambda_2) + \lambda_2 = -u_{2,2}\nu_2 \Rightarrow u_{2,2} = -\frac{1}{\nu_2}[d_1(\lambda_1 - \lambda_2) + \lambda_2].$$

We are left with one polynomial equation of degree 2 and one unknown d_1,

$$d_1(\lambda_1^2 - \lambda_2^2) + \lambda_2^2 = \frac{1 + \nu_2^2}{\nu_2^2}[d_1(\lambda_1 - \lambda_2) + \lambda_2]^2.$$

After some computations, it turns out that the discriminant is

$$\frac{1 + \nu_2^2}{\nu_2^2}(\lambda_1 - \lambda_2)^2[-4\lambda_1\lambda_2 + \frac{1 + \nu_2^2}{\nu_2^2}(\lambda_1 + \lambda_2)^2].$$

Clearly this can be negative if ν_2^2 is small enough that is, if

$$\nu_2^2 < \frac{4\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)^2}.$$

If this is the case there is no real solution to the equation and consequently we cannot construct U and the matrix T with prescribed eigenvalues. Even, if a real solution exists, there are constraints on the eigenvalues for the d_j's to be positive. Hence, contrary to the nonsymmetric case, residual norms and eigenvalues cannot always be prescribed together. This does not come as a surprise since it is well known that CG convergence depends on the distribution of the eigenvalues.

8 Numerical experiments

Let us illustrate the results of the previous sections with four small examples. The entries of T were computed using formulas (11) and (19). In this section we use the word "A-norm" but for three of the examples we have $A = T$.

8.1 Example 1

For the first example, we construct a linear system for which the CG residual norms are not converging before the last iteration but the A-norms of the error are decreasing fast. We choose $n = 15$ and the prescribed residual norms as

$$f_0 = 1, \quad f_{2i-1} = 2, \quad f_{2i} = 1,\ i = 1, \ldots$$

Hence the residual norms oscillate, being 1 or 2 depending on the parity of the iteration number. The values of the A-norm of the error are $g_0 = 1, \quad g_i = 0.6\, g_{i-1},\ i = 1, \ldots, n$. The condition number of the constructed matrix T is $6.37 \cdot 10^6$. We solve $Tx = e_1$ with CG. Figure 1 shows the prescribed and computed residual norms

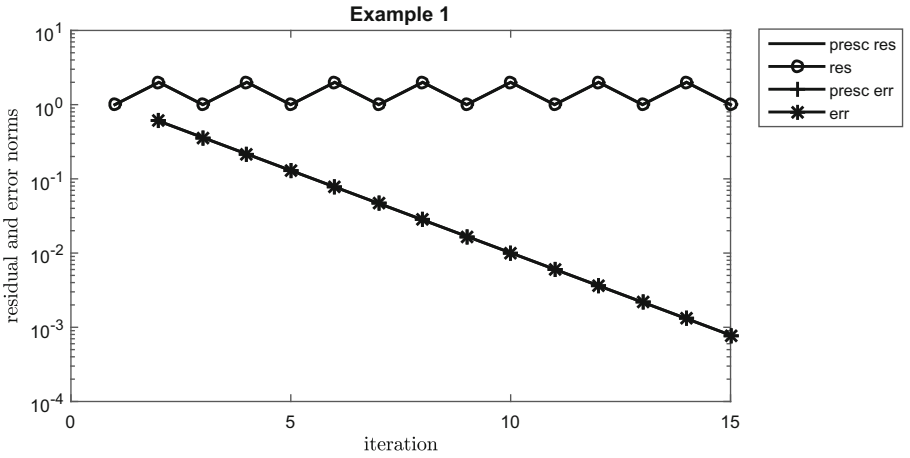

Fig. 1 Example 1, the prescribed and computed residual norms and A-norms of the error

and A-norms of the error. The prescribed and computed values are visually indistinguishable. Figure 2 displays their relative differences. We can see that they are at the roundoff level. The missing values are exactly zero.

8.2 Example 2

In the second example we take $n = 20$ and the prescribed residual norms are [1; 0.9; 0.8; 0.6; 0.3; 0.1; 0.09; 0.08; 0.06; 0.03; 0.01; 0.009; 0.008; 0.006; 0.003; 0.001; 0.0005; 0.0001; 0.00005; 0.00001]. The A-norms of the error are defined as $g_0 = 1, \quad g_i = 0.3\, g_{i-1},\ i = 1, \ldots, n$. The condition number of the matrix T is $2.20{\cdot}10^{10}$.

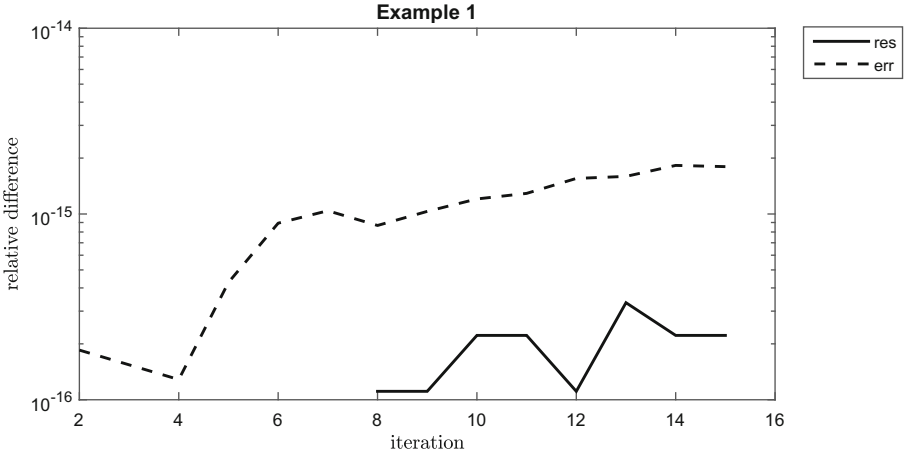

Fig. 2 Example 1, relative differences of the residual norms and A-norms of the error with the prescribed ones

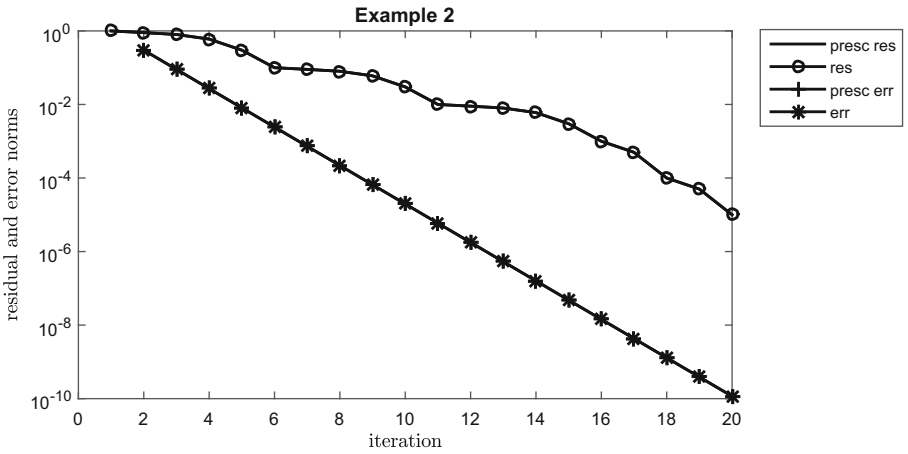

Fig. 3 Example 2, the prescribed and computed residual norms and A-norms of the error

Figure 3 shows the prescribed and computed residual norms and A-norms of the error. Again, the prescribed and computed values are visually indistinguishable. Figure 4 displays their relative differences. We can check that they are at the roundoff level except for the last iterations for the A-norm.

8.3 Example 3

We use the same order and prescribed residual norms as in example 2. In this example we use a matrix $A = VTV^T$ and a right-hand side $b = Ve_1$ where V is a random orthonormal matrix. The A-norms of the error are defined as 100 times the residual norms. The condition number of the matrix A is 61.6.

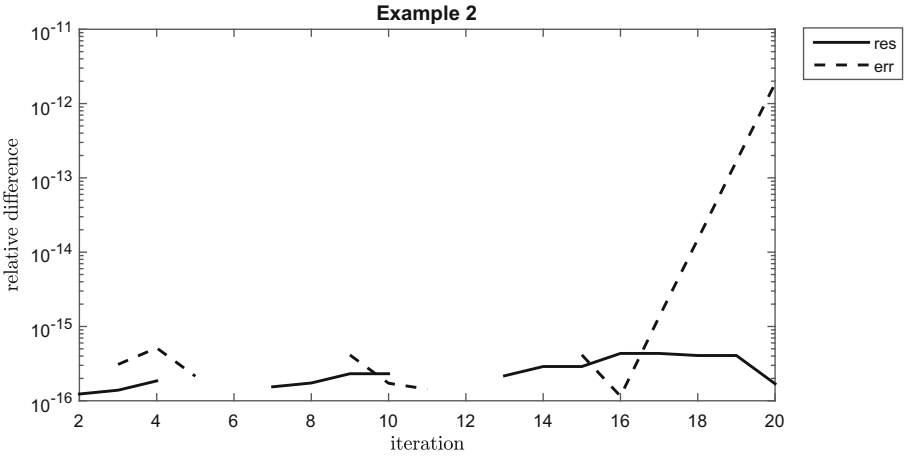

Fig. 4 Example 2, relative differences of the residual norms and A-norms of the error with the prescribed ones

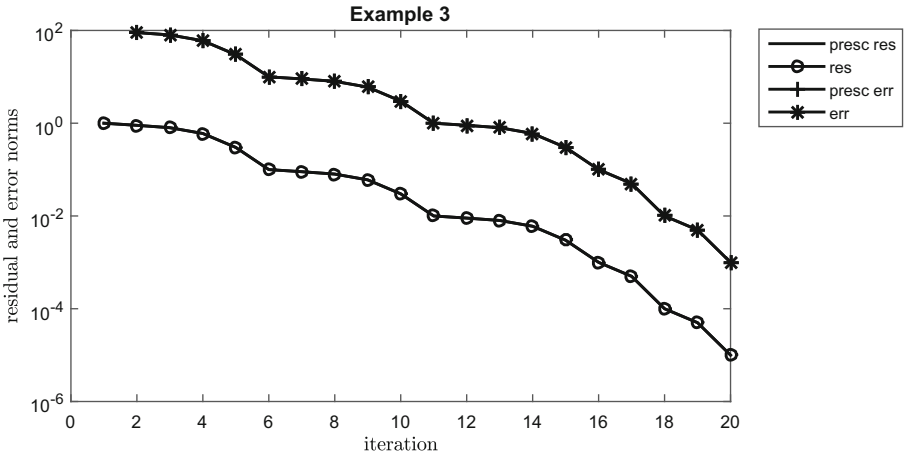

Fig. 5 Example 3, the prescribed and computed residual norms and A-norms of the error

Figure 5 shows the prescribed and computed residual norms and A-norms of the error. Again, the prescribed and computed values are visually indistinguishable. Figure 6 displays their relative differences. We can see that they are slightly larger than the roundoff level except for the last iterations.

8.4 Example 4

For this example we just prescribe the residual norms as $f_0 = 1$, $f_i = 0.8\, f_{i-1}$, $i = 1, \ldots, n$ with $n = 20$ and $\sigma_i \equiv 1$. The condition number of the matrix T is $2.82 \cdot 10^3$. In this example, in finite precision arithmetic, CG suffers from

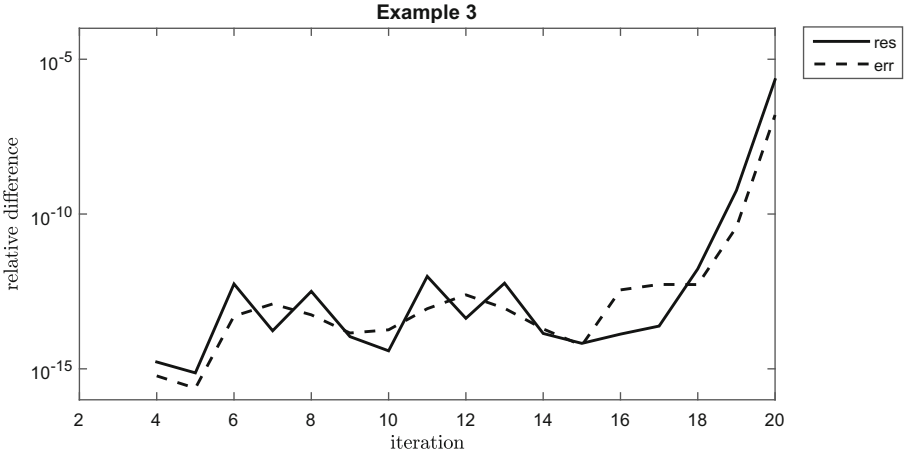

Fig. 6 Example 3, relative differences of the residual norms and A-norms of the error with the prescribed ones

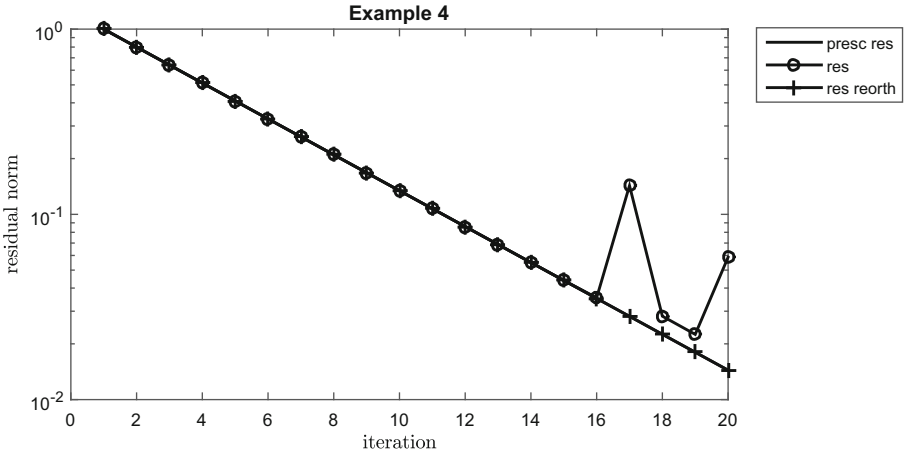

Fig. 7 Example 4, the prescribed and computed residual norms with and without reorthogonalization

rounding errors and residual vectors lose their orthogonality (see [18, 20]). This leads to a delay in convergence. We can see in Fig. 7 that the computed residual norms deviate from the prescribed values. However, if we reorthogonalize the residual vectors, we obtain the prescribed values. This can also be seen in Fig. 8 which displays the relative differences. What we have prescribed is the mathematical behavior of CG, not the computed one. However, our construction can be used to study the differences between the mathematical properties and the behavior in finite precision arithmetic.

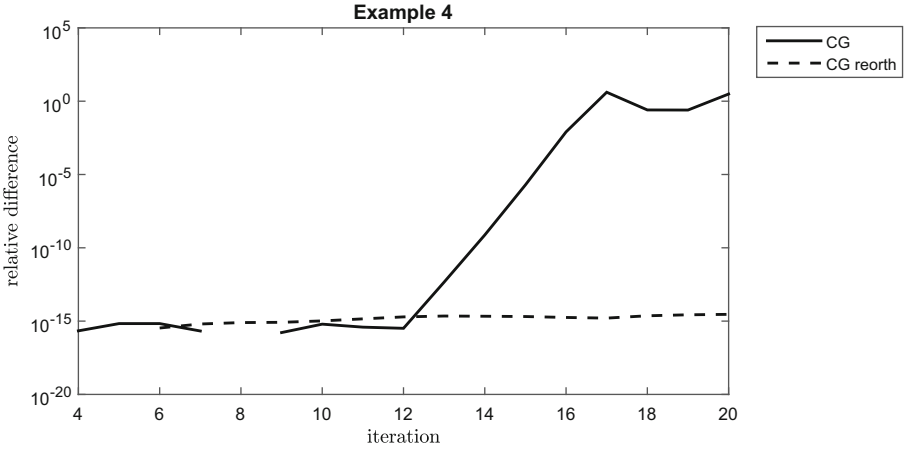

Fig. 8 Example 4, relative differences of the residual norms with the prescribed ones with and without reorthogonalization

9 Conclusion

In this paper we have shown how to construct linear systems with a positive definite symmetric matrix such that CG gives prescribed residual norms as well as prescribed decreasing A-norms of the error. It means that any convergence situation can happen for CG. The difference with the situation for nonsymmetric problems and FOM/GMRES (see [1, 4, 9, 10]) is that it is not always possible to also prescribe the eigenvalues of the matrix. But, this does not come as a surprise since CG convergence depends on the distribution of the eigenvalues of the matrix (see, for instance, [18]).

Finally, we observe that, since there are relations between the residual norms in CG and those in Minres (see [21]), we can also prescribe the residual norms for Minres.

Acknowledgments Many thanks to Erin Carson and Jurjen Duintjer Tebbens for interesting comments and suggestions and to Petr Tichý for comments and suggesting to use relation (2). The author thanks the referees for their detailed comments.

References

1. Arioli, M., Pták, V., Strakoš, Z.: Krylov sequences of maximal length and convergence of GMRES. BIT Numer. Math. **38**(4), 636–643 (1998)
2. Arnoldi, W.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. Quart. Appl. Math. **9**, 17–29 (1951)
3. Baranger, J., Duc-Jacquet, M.: Matrices tridiagonales symétriques et matrices factorisables. RIRO **3**, 61–66 (1971)
4. Duintjer Tebbens, J., Meurant, G.: Any Ritz value behavior is possible for Arnoldi and for GMRES. SIAM J. Matrix Anal. Appl. **33**(3), 958–978 (2012)
5. Duintjer Tebbens, J., Meurant, G.: Prescribing the behavior of early terminating GMRES and Arnoldi iterations. Numer. Algorithms **65**(1), 69–90 (2014)
6. Duintjer Tebbens, J., Meurant, G.: On the convergence of Q-OR and Q-MR Krylov methods for solving linear systems. BIT Numer. Math. **56**(1), 77–97 (2016)
7. Fiedler, M.: Polynomials and Hankel matrices. Linear Algebra Appl. **66**, 235–248 (1985)
8. Golub, G.H., Meurant, G.: Matrices, Moments and Quadrature with Applications. Princeton University Press (2010)
9. Greenbaum, A., Strakoš, Z.: Matrices that generate the same Krylov residual spaces. In: Golub, G.H., Greenbaum, A., Luskin, M. (eds.) Recent Advances in Iterative Methods, pp. 95–118. Springer (1994)
10. Greenbaum, A., Pták, V., Strakoš, Z.: Any convergence curve is possible for GMRES. SIAM J. Matrix Anal. Appl. **17**(3), 465–470 (1996)
11. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Nat. Bur. Standards **49**(6), 409–436 (1952)
12. Ikebe, Y.: On inverses of Hessenberg matrices. Linear Algebra Appl. **24**, 93–97 (1979)
13. Juárez-Ruiz, E., Cortés-Maldonado, R., Pérez-Rodríguez, F.: Relationship between the inverses of a matrix and a submatrix. Computación y Sistemas **20**(2), 251–262 (2016)
14. Lanczos, C.: Solution of systems of linear equations by minimized iterations. J. Res. Nat. Bur. Standards **49**, 33–53 (1952)
15. Liesen, J., Strakoš, Z.: Krylov Subspace Methods, Principles and Analysis. Oxford University Press (2013)
16. Meurant, G.: A review on the inverse of tridiagonal and block tridiagonal symmetric matrices. SIAM J. Matrix Anal. Appl. **13**(3), 707–728 (1992)
17. Meurant, G.: Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. Numer. Algorithms **22**, 353–365 (1999)

18. Meurant, G.: The Lanczos and Conjugate Gradient Algorithms from Theory to Finite Precision Computations. SIAM, Philadelphia (2006)
19. Meurant, G., Strakoš, Z.: The Lanczos and conjugate gradient algorithms in finite precision arithmetic. Acta Numerica **15**, 471–542 (2006)
20. Paige, C.C.: The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices. University of London, Ph.D. thesis (1971)
21. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12**(4), 617–629 (1975)
22. Saad, Y.: Krylov subspace methods for solving large nonsymmetric linear systems. Math. Comp. **37**, 105–126 (1981)
23. Saad, Y.: Practical use of some Krylov subspace methods for solving indefinite and nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. **5**(1), 203–228 (1984)
24. Schweitzer, M.: Any finite convergence curve is possible in the initial iterations of restarted FOM. Electron. Trans. Numer. Anal. **45**, 133–145 (2016)
25. Strakoš, Z., Tichý, P.: On error estimates in the conjugate gradient method and why it works in finite precision computations. Electron. Trans. Numer. Anal. **13**, 56–80 (2002)