On prescribing CG convergence

Gérard MEURANT

M2A19, Rabat, April 2019

The Conjugate Gradient (CG) is the algorithm of choice for solving iteratively linear systems

$$Ax = b$$

when A is symmetric and positive definite (SPD) of order n

M.R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Nat. Bur. Standards, v 49 n 6 (1952), pp. 409–436

This paper is still worth reading!

Conjugate Gradient (CG)

input A, b, x_0

$r_0 = b - Ax_0$

$p_0 = r_0$

for $k = 1, \ldots$ until convergence **do**

$\sigma_{k-1} = \frac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$

$x_k = x_{k-1} + \sigma_{k-1} p_{k-1}$

$r_k = r_{k-1} - \sigma_{k-1} A p_{k-1}$

$\delta_k = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$

$p_k = r_k + \delta_k p_{k-1}$

end for

CG minimizes the A-norm of the error

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

which is decreasing

We would like to see if we can construct matrices A and right-hand sides b such that the residual norms $\|r_k\|$ have prescribed positive values

The answer to this question was (almost) given by Hestenes and Stiefel page 432

derivative of the characteristic polynomial of A into a continued fraction.

Finally, we are able to prove

Theorem 18:3. *There is no restriction whatever on the positive constants a_i, b_i in the cg-process, that is, given two sequences of positive numbers $a_0, a_1, \ldots, a_{n-1}$ and $b_0, b_1, \ldots, b_{n-1}$, 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.*

The demonstration goes along the following lines: From (15:2) and (15:4), we compute the numbers c_i, d_i, the c_i being again positive. Then we use the continued fraction (18:2) to compute $F(\lambda)$ which we decompose into partial fractions to obtain (18:1). We show next that the numbers λ_i, m_i appearing in (18:1) are positive. After this has been established, our correspondence finishes the proof.

In order to prove that $\lambda_i > 0$, $m_i > 0$ we observe that the ratio R_{i+1}/R_i is a decreasing function of λ, as can be seen from (18:3) by induction. Using this result, it is not too difficult to show that the polynomials $R_0(\lambda)$, $R_1(\lambda)$, $\ldots$, $R_n(\lambda)$ build a Sturmian sequence in the following sense. The number of zeros of $R_n(\lambda)$ in any interval $a \leq \lambda \leq b$ is equal to the increase of the number of variations in sign in going from a to b. At the point λ_0 there are no variations in sign since $R_i(0) = 1$ for every i. At $\lambda = -\infty$, there are exactly n variations because the coefficient of the highest power of λ in $R_i(\lambda)$ is $(-1)^i a_0 a_1 \ldots a_{i-1}$. Therefore, the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $R_n(\lambda)$ are real and positive.

That the function $F(\lambda)$ is itself a decreasing function of λ follows directly from (18:2). Therefore, its residues $m_1, m_2, \ldots, m_n$ are positive.

In view of theorem 18:3 the numbers a_i in a cg-process can increase as fast as desired. This result was used in section 8.2. 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. Hence, there are certainly examples where the residual vectors increase in length during the computation, as was stated earlier. This holds in spite of the fact that the error vector $h - x$ decreases in length at each step.

If you were not able to read that page. . .

Theorem 18.3

There is no restriction whatever on the positive constants a_i, b_i (*our* σ_k *and* δ_k) 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 . . .

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

Hence, it seems we are done!

However, we will show that we can construct matrices and right-hand sides such that we can prescribe the residual norms and the A-norms of the error

CG is related to the Lanczos algorithm. It implicitly constructs a Cholesky factorization of the Lanczos tridiagonal matrix

We will construct symmetric tridiagonal matrices T such that CG (with $x_0 = 0$) yields prescribed relative residual norms with the right-hand side e_1
Then, the linear system $Ax = b$ is obtained from $A = VTV^T$ and $b = Ve_1$ where V is any orthonormal matrix

Strangely enough, let us start with the nonsymmetric case...

The nonsymmetric case

We consider the Full Orthogonalization Method (FOM) of Y. Saad for solving nonsymmetric linear systems

$$\text{FOM} \equiv \text{Lanczos} \equiv \text{CG} \text{ when } A \text{ is symmetric}$$

Prescribing the residual norms for FOM was studied some years ago

We can construct linear systems for which the matrix has prescribed eigenvalues and such that FOM delivers prescribed residual norms and gives also prescribed Ritz values at all iterations, see

J. Duintjer Tebbens and G. Meurant, *Any Ritz value behavior is possible for Arnoldi and for GMRES*, SIAM J. Matrix Anal. Appl., v 33, n 3 (2012), pp. 958–978

Assume FOM does not stop early, then we get

$$AV = VH$$

V is the orthonormal matrix whose columns are the Arnoldi basis vectors and H is an unreduced upper Hessenberg matrix with positive subdiagonal entries

$$H = UCU^{-1}$$

where U is upper triangular with $u_{1,1} = 1$ and C is the companion matrix corresponding to the eigenvalues of A (and H)

$$U = \begin{pmatrix} e_1 & He_1 & \cdots & H^{n-1}e_1 \end{pmatrix}$$

It was proved (see JDT-GM) 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 $\|r_k^F\|/\|r_0\|$

Note that the other entries of U^{-1} can be chosen to prescribe the Ritz values but we won't use that in the symmetric case

1) We construct an upper triangular matrix U^{-1} with appropriate values on the first row

2) With the prescribed eigenvalues, we construct the companion matrix C

3) $H = UCU^{-1}$

4) $A = VHV^T$, $b = Ve_1$, V being any orthonormal matrix

However, it is difficult to construct a symmetric H in this way

So, we proceed differently...

Now we use results in
A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent advances in iterative methods, G.H. Golub, A. Greenbaum and M. Luskin Eds., Springer (1994), pp. 95–118

Theorem

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 $k = n$ and $Y = RH$

Then $R^{-1} = HY^{-1}$ and the matrix HY^{-1} must be upper triangular

From G and S, HX is upper triangular if and only if

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

where $\mathcal{L}(F)$ is 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

It 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 $H = UCU^{-1}$ for which FOM applied to (H, e_1) gives the prescribed residual norms

We are looking for a nonsingular matrix X for which FOM gives the same residual norms on (X^{-1}, e_1)

It is sufficient to construct a lower triangular matrix X whose lower triangular part is the same as that of H^{-1}

Let $H = UCU^{-1}$ with

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

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

Let $\hat{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T$

Since $\alpha_0 \neq 0$, C is nonsingular and its inverse is

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

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,

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

$$H^{-1} = UC^{-1}U^{-1}$$

Theorem

Using the previous notation the inverse of H is

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

The matrix $\hat{U} E \hat{U}^{-1}$ is strictly upper triangular. Hence, we have a simple proof of a well-known result about inverses of Hessenberg matrices

Namely, the lower triangular part of H^{-1} is the same as the lower triangular part of a rank-one matrix

Y. Ikebe, *On inverses of Hessenberg matrices*, Linear Algebra Appl., v 24 (1979), pp. 93–97

Let us construct a lower triangular matrix X such that $\mathcal{L}(X) = \mathcal{L}(H^{-1})$

To simplify the notation, let $\tilde{\beta} = \beta_1 - \hat{\nu}^T \hat{U} \hat{\beta}$

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

with $\tilde{U} = \mathcal{L}(\hat{U}\hat{\beta}\hat{\nu}^T)$

The matrix X^{-1} is, in fact, lower bidiagonal

Theorem

Let $\gamma = \hat{U}\hat{\beta}$, D_ν *(resp.* D_γ*) be the diagonal matrix whose diagonal entries are* $1/\nu_i$, $i = 1, \ldots, n$ *(resp.* $1/\gamma_{i-1}$ *with* $\gamma_0 = \tilde{\beta}$*) and* B *be the bidiagonal matrix*

$$B = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix}$$

Then, $X^{-1} = D_\nu B D_\gamma$

$$X = D_\gamma^{-1} B^{-1} D_\nu^{-1} = \begin{pmatrix} \gamma_0 & & & \\ \gamma_1 & \gamma_1\nu_2 & & \\ \vdots & \vdots & \ddots & \\ \gamma_{n-1} & \gamma_{n-1}\nu_2 & \cdots & \gamma_{n-1}\nu_n \end{pmatrix}$$

In this matrix, the ν_i's are linked to the FOM residual norms and the γ_i's appear as parameters that we are free to choose

We can prove that FOM applied to (X^{-1}, e_1) yields residual norms such that $\|r_0\|/\|r_k^F\| = \nu_{k+1}$

The symmetric case

We follow G-S and define

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

where $\hat{\mathcal{L}}$ gives the strict lower triangular part of the matrix

$$X_s = \begin{pmatrix} \tilde{\beta} & (\hat{U}\hat{\beta})^T \\ \hat{U}\hat{\beta} & \mathcal{L}(\hat{U}\hat{\beta}\hat{\nu}^T) + \hat{\mathcal{L}}(\hat{U}\hat{\beta}\hat{\nu}^T)^T \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_1\nu_2 & \cdots & \gamma_{n-1}\nu_2 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-1}\nu_2 & \cdots & \gamma_{n-1}\nu_n \end{pmatrix}$$

From the structure of X_s we know that X_s^{-1} is a symmetric tridiagonal matrix whose nonzero entries can be computed from the entries of X_s

Let $\nu = \begin{pmatrix} 1 & \nu_2 & \cdots & \nu_n \end{pmatrix}^T$ and μ_i, $i = 1, \ldots, n$, η_i, $i = 1, \ldots, n-1$ be the diagonal and subdiagonal entries of the tridiagonal matrix X_s^{-1}

$$\mu_1 = -\frac{\nu_2}{\gamma_1 - \nu_2\gamma_0}, \quad \eta_1 = \frac{1}{\gamma_1 - \nu_2\gamma_0} = -\frac{\mu_1}{\nu_2},$$

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

$$d_i = \nu_i(\nu_i\gamma_i - \nu_{i+1}\gamma_{i-1})$$

$$\mu_i = -\frac{\nu_{i+1}}{d_i} - \eta_{i-1}\frac{\nu_{i-1}}{\nu_i}, \quad \eta_i = \frac{\nu_i}{d_i},$$

the last diagonal entry being equal to

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

We have to assume that $d_i \neq 0$

We claim that FOM ($\equiv$ CG) applied to (X_s^{-1}, e_1) gives relative residual norms equal to the inverses of the absolute values of the components of ν

Let $U_{X_s^{-1}} = \begin{pmatrix} e_1 & X_s^{-1}e_1 & \cdots & X_s^{-(n-1)}e_1 \end{pmatrix}$

We would like to show that $e_1^T U_{X_s^{-1}}^{-1} = \nu^T$ that is, $\nu^T U_{X_s^{-1}} = e_1^T$

We can first prove that $\nu^T X_s^{-1} e_i = 0$ for $i = 1, \ldots, n-1$

Only the last component of $\nu^T X_s^{-1}$ is nonzero and the $(n, 1)$ entry of the i-th power of X_s^{-1} is zero for $j \leq n-2$ and it follows that $\nu^T U_{X_s^{-1}} = e_1^T$

However for CG we need the tridiagonal matrix to be positive definite

We can choose the γ_i's to obtain a positive definite matrix

Let us compute the Cholesky-like factorization $X_s^{-1} = L\Omega^{-1}L^T$ with L lower bidiagonal and Ω diagonal. Let ω_i, $i = 1, \ldots, n$ be the diagonal entries of L and Ω

$$\omega_i = -\frac{\nu_{i+1}}{\nu_i} \frac{1}{\nu_i\gamma_i - \nu_{i+1}\gamma_{i-1}}, \ i = 1, \ldots, n-1, \quad \omega_n = \frac{1}{\nu_n\gamma_{n-1}}$$

We have $\omega_i = 1/\sigma_{i-1}$. If $\omega_i > 0$, the matrix $T = X_s^{-1}$ is positive definite

Note that γ_{i-1} and ν_i have to be of the same sign for T^{-1} to be positive definite

If we choose the ν_i's and γ_i's strictly positive, the condition is

$$\frac{\gamma_{i-1}}{\gamma_i} > \frac{\nu_i}{\nu_{i+1}}, \ i = 1, \ldots, n-1$$

Prescribing the A-norm of the error

Let $\varepsilon_k = x - x_k$ be the error vector, $A = VTV^T$. It is known that

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

where T_k is the principal submatrix of T of order k

See, for instance, Theorem 12.1 in
G.H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*, Princeton University Press, (2010)

We have to link the entries of T_k^{-1} to those of T^{-1}

This can be done in different ways, by induction using the Cholesky factorization or more directly. The result is the following

Theorem

$$(T^{-1})_{1,1} - (T_k^{-1})_{1,1} = \frac{\gamma_k}{\nu_{k+1}}, \ k = 1, \ldots, n-1$$

Corollary

The square of the A-norm of the error is given by

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

If we prescribe decreasing values for $\|\varepsilon_k\|_A$ we obtain the values of the γ_k's

Prescribing decreasing values corresponds to the condition to have T positive definite

In fact, the inverse of T_k is

$$T_k^{-1} = \begin{pmatrix} \gamma_0^{(k)} & \gamma_1^{(k)} & \cdots & \gamma_{k-1}^{(k)} \\ \gamma_1^{(k)} & \gamma_1^{(k)}\nu_2 & \cdots & \gamma_{k-1}^{(k)}\nu_2 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k-1}^{(k)} & \gamma_{k-1}^{(k)}\nu_2 & \cdots & \gamma_{k-1}^{(k)}\nu_k \end{pmatrix}$$

with

$$\gamma_i^{(k)} = \gamma_i^{(k+1)} - \gamma_k^{(k+1)} \frac{\nu_{i+1}}{\nu_{k+1}}, \ i = 0, 1, \ldots, k-1$$

Moreover,

$$\gamma_i^{(k)} = \gamma_i - \gamma_k \frac{\nu_{i+1}}{\nu_{k+1}}, \ i = 0, 1, \ldots, k-1$$

It yields $\forall i = 0, \ldots, k-1$

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

Mathematically, any residual norm convergence curve is possible for CG with any decreasing A-norm convergence curve

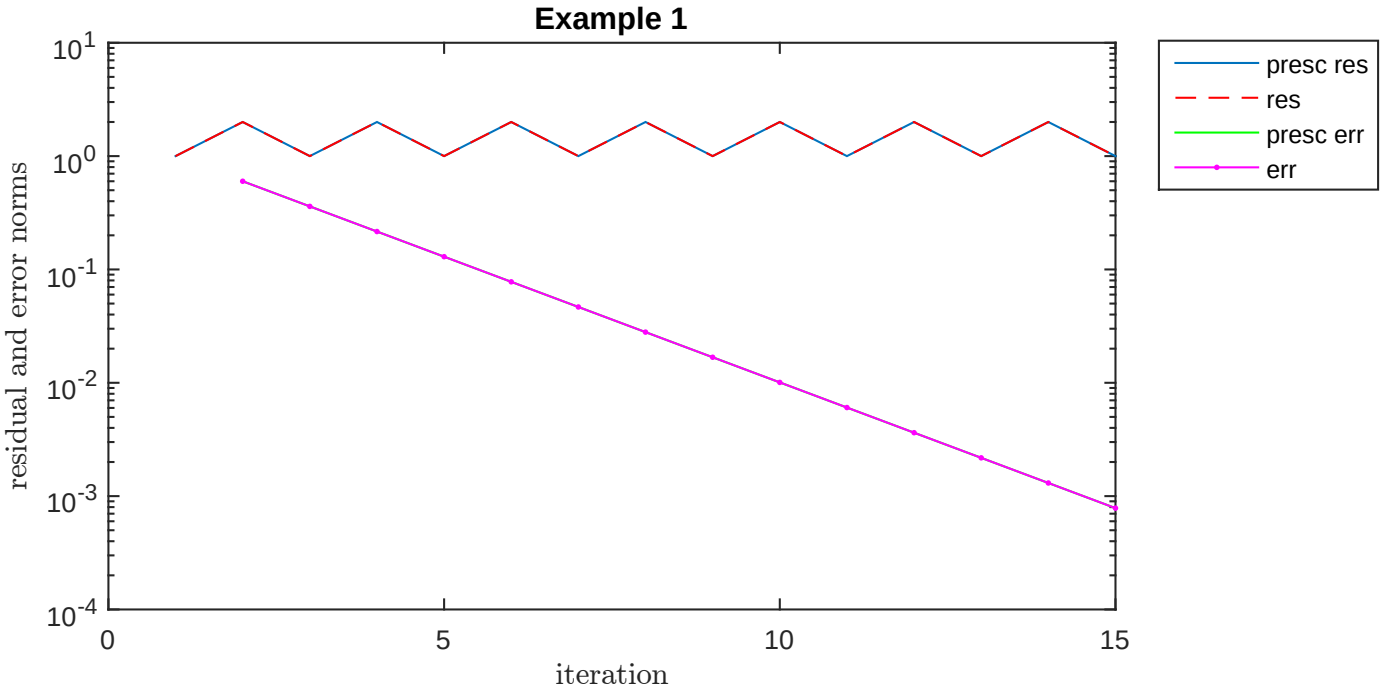
Let us consider some numerical experiments

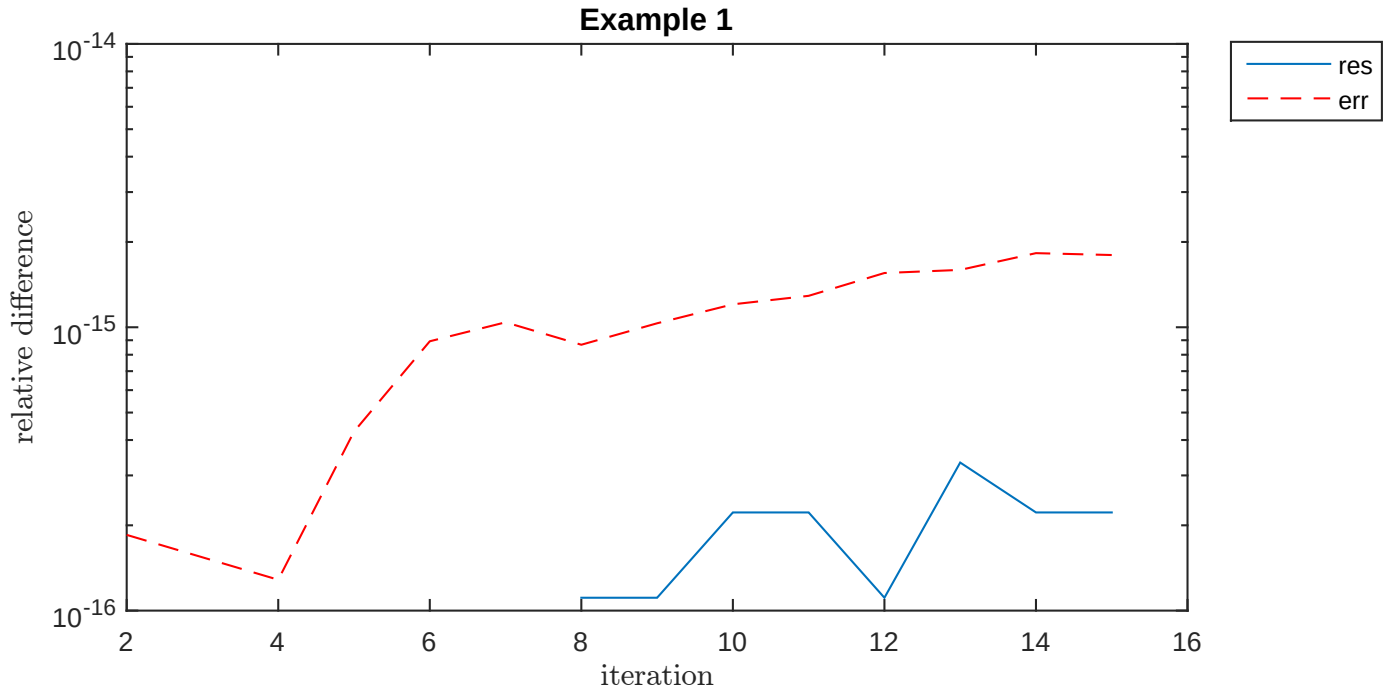
Numerical experiments

Example 1

$n = 15$, res $= [1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1]$, err:

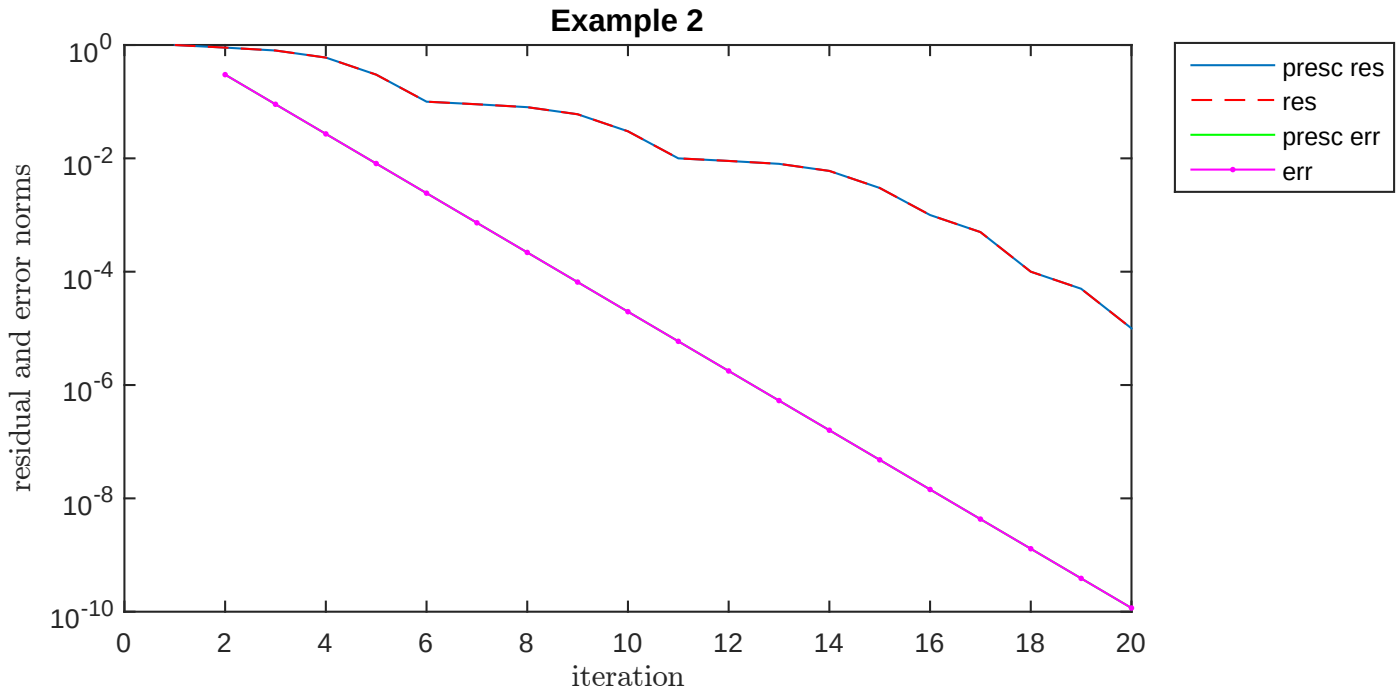
$g_0 = 1, \quad g_i = 0.6\, g_{i-1},\ i = 1, \ldots, n$, cond$(T) = 6.37\ 10^6$

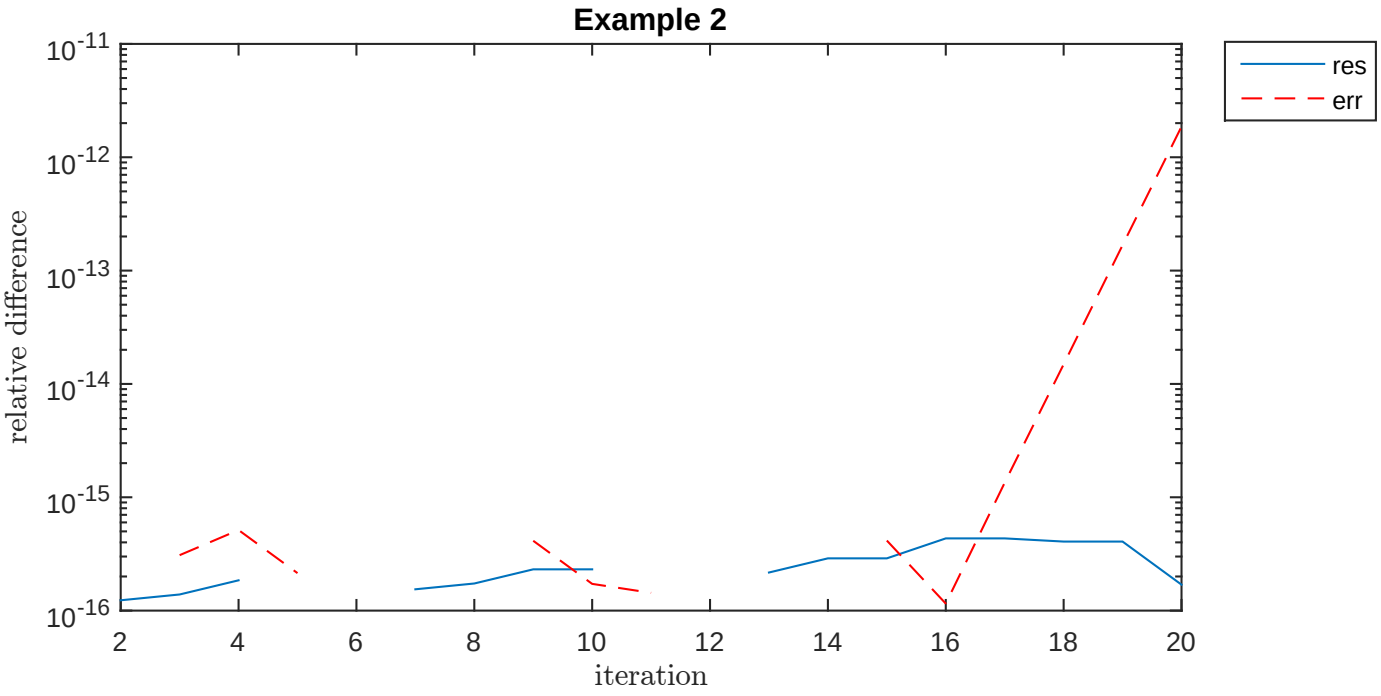



Example 1
relative difference
iteration
res
err
10^-14
10^-15
10^-16
2
4
6
8
10
12
14
16

Example 2

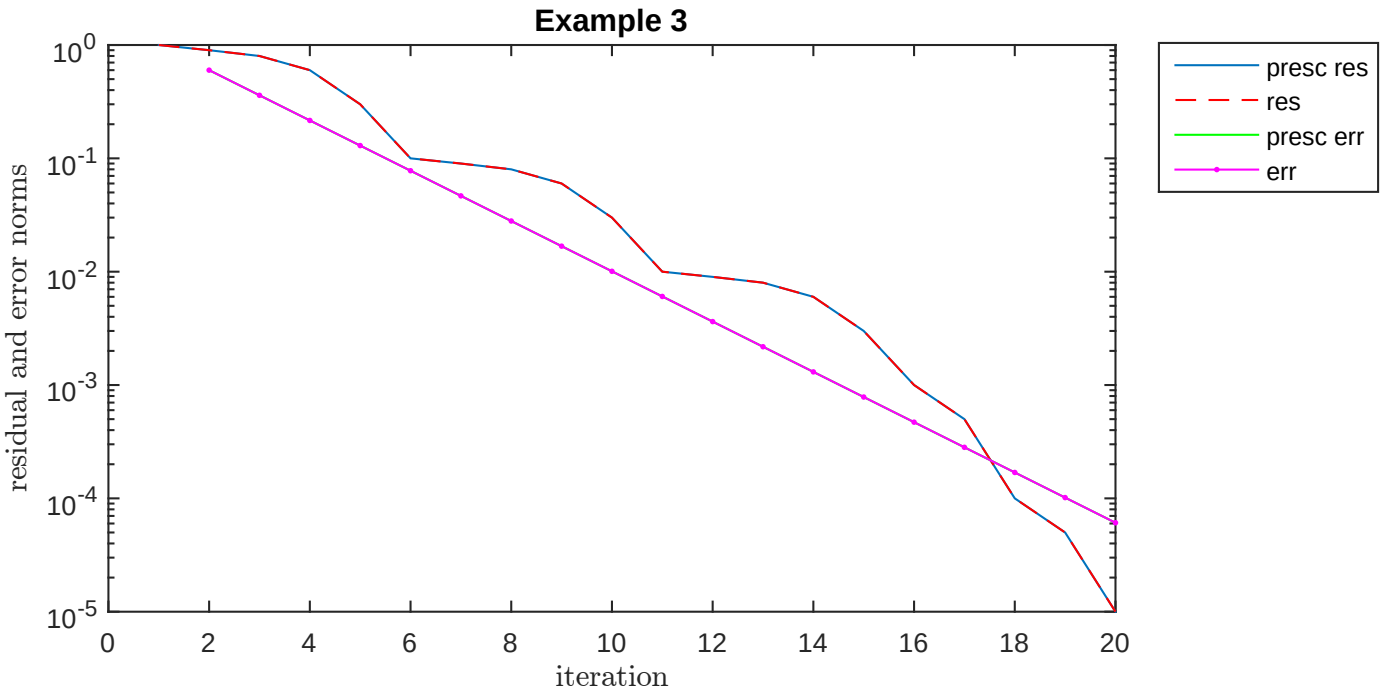
$n = 20$, res $= [1; 0.9; 0.8; 0.6; 0.3; 0.1; 0.09; 0.08; 0.06; 0.03; 0.01; 0.009; 0.008; 0.006; 0.003; 0.001; 0.0005; 0.0001; 0.00005; 0.00001]$, err: $g_0 = 1, \quad g_i = 0.3\, g_{i-1}, \ i = 1, \ldots, n$, cond$(T) = 2.20\ 10^{10}$

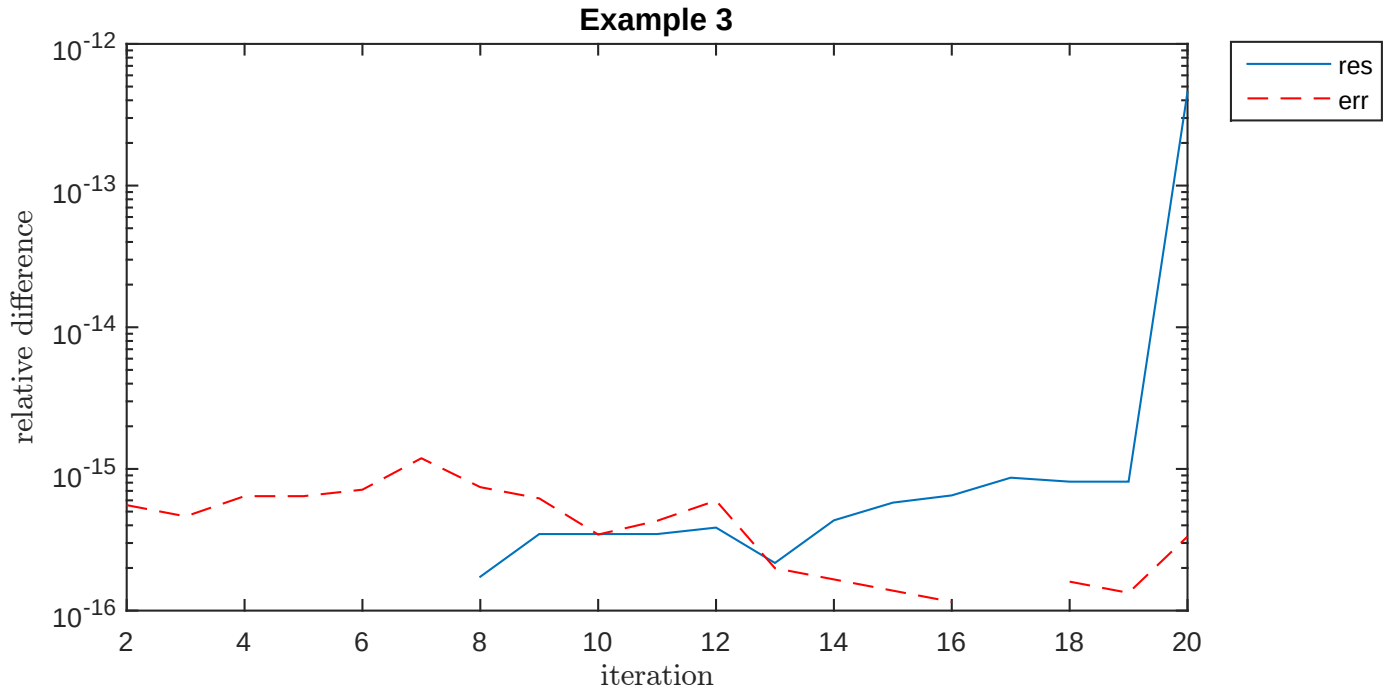



Example 2
relative difference
iteration
res
err
10^-11
10^-12
10^-13
10^-14
10^-15
10^-16
2
4
6
8
10
12
14
16
18
20

Example 3

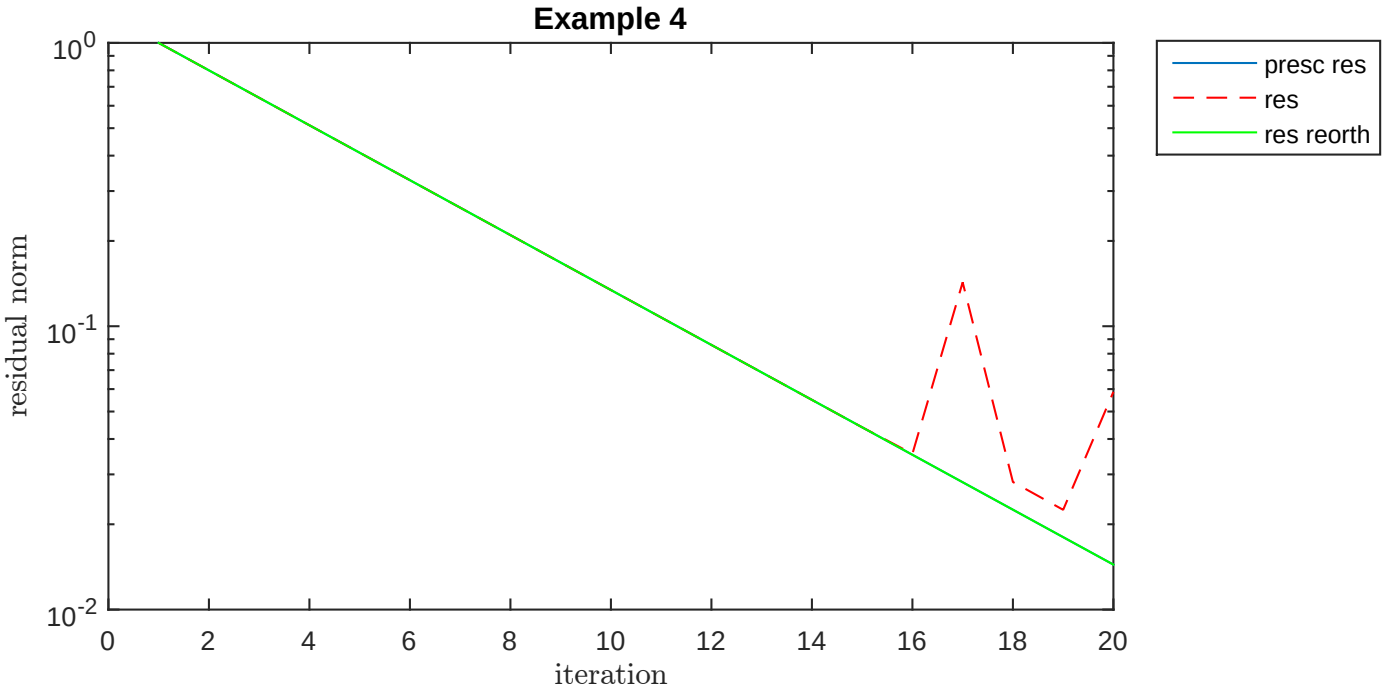
$n = 20$, res $= [1; 0.9; 0.8; 0.6; 0.3; 0.1; 0.09; 0.08; 0.06; 0.03; 0.01; 0.009; 0.008; 0.006; 0.003; 0.001; 0.0005; 0.0001; 0.00005; 0.00001]$, err: $g_0 = 1, \quad g_i = 0.6\, g_{i-1}, \ i = 1, \ldots, n$, cond$(T) = 2.43\ 10^3$



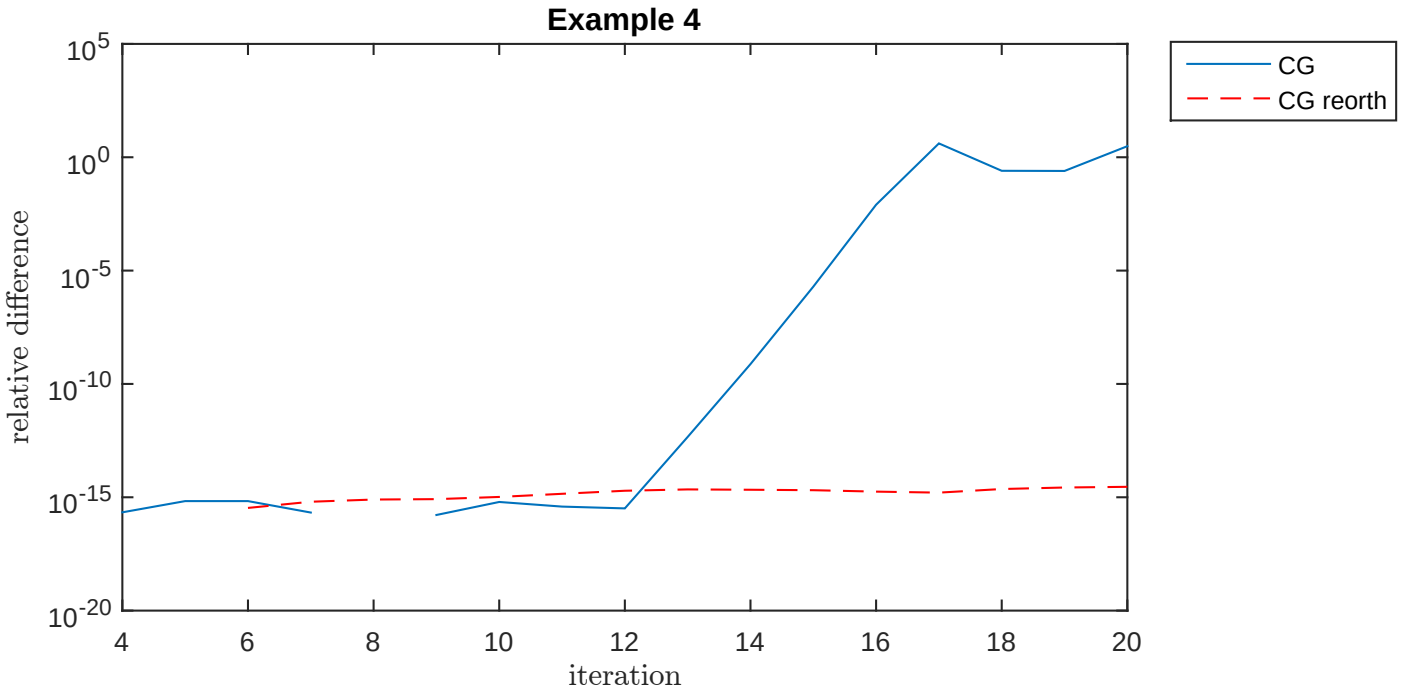

Example 3
relative difference
iteration
res
err
10^-12
10^-13
10^-14
10^-15
10^-16
2
4
6
8
10
12
14
16
18
20

Example 4

$n = 20$, res: $f_0 = 1, \quad f_i = 0.8\, f_{i-1}, \ i = 1, \ldots, n$, err: $\gamma_i \equiv 1$,
$\mathrm{cond}(T) = 2.82\ 10^3$



In this example, in finite precision arithmetic, CG suffers from rounding errors and residual vectors loose their orthogonality


Example 4
relative difference
iteration
CG
CG reorth
10^5
10^0
10^-5
10^-10
10^-15
10^-20
4
6
8
10
12
14
16
18
20

Conclusion

We have shown how to construct symmetric positive definite linear systems with a prescribed CG residual norm convergence curve as well as a prescribed A-norm of the error convergence curve