ORIGINAL PAPER

The behavior of the Gauss-Radau upper bound of the error norm in CG

Gérard Meurant[1] · Petr Tichý[2]

Abstract

Consider the problem of solving systems of linear algebraic equations $Ax = b$ with a real symmetric positive definite matrix A using the conjugate gradient (CG) method. To stop the algorithm at the appropriate moment, it is important to monitor the quality of the approximate solution. One of the most relevant quantities for measuring the quality of the approximate solution is the A-norm of the error. This quantity cannot be easily computed; however, it can be estimated. In this paper we discuss and analyze the behavior of the Gauss-Radau upper bound on the A-norm of the error, based on viewing CG as a procedure for approximating a certain Riemann-Stieltjes integral. This upper bound depends on a prescribed underestimate μ to the smallest eigenvalue of A. We concentrate on explaining a phenomenon observed during computations showing that, in later CG iterations, the upper bound loses its accuracy, and is almost independent of μ. We construct a model problem that is used to demonstrate and study the behavior of the upper bound in dependence of μ, and developed formulas that are helpful in understanding this behavior. We show that the above-mentioned phenomenon is closely related to the convergence of the smallest Ritz value to the smallest eigenvalue of A. It occurs when the smallest Ritz value is a better approximation to the smallest eigenvalue than the prescribed underestimate μ. We also suggest an adaptive strategy for improving the accuracy of the upper bounds in the previous iterations.

✉ Petr Tichý
petr.tichy@mff.cuni.cz

Gérard Meurant
gerard.meurant@gmail.com

[1] Paris, France

[2] Faculty of Mathematics and Physics, Charles University, Sokolovská 83, Prague 18675, Czech Republic

Ⓓ Springer

1 Introduction

Our aim in this paper is to explain the origin of the problems that have been noticed [1] when computing Gauss-Radau quadrature upper bounds of the A-norm of the error in the Conjugate Gradient (CG) algorithm for solving linear systems $Ax = b$ with a symmetric positive definite matrix of order N.

The connection between CG and Gauss quadrature has been known since the seminal paper of Hestenes and Stiefel [2] in 1952. This link has been exploited by Gene H. Golub and his collaborators to bound or estimate the A-norm of the error in CG during the iterations; see [1, 3–13].

Using a fixed node μ smaller than the smallest eigenvalue of A and the Gauss-Radau quadrature rule, an upper bound for the A-norm of the error can be easily computed. Note that it is useful to have an upper bound of the error norm to stop the CG iterations. In theory, the closer μ is to the smallest eigenvalue, the closer is the bound to the norm.

Concerning the approximation properties of the upper bound, we observed in many examples that in earlier iterations, the bound is approximating the A-norm of the error quite well, and that the quality of approximation is improving with increasing iterations. However, in later CG iterations, the bound suddenly becomes worse: it is delayed, almost independent of μ, and does not represent a good approximation to the A-norm of the error any more. Such a behavior of the upper bound can be observed also in exact arithmetic. Therefore, the problem of the loss of accuracy of the upper bound in later iterations is not directly linked to rounding errors and has to be explained.

The Gauss quadrature bounds of the error norm were obtained by using the connection of CG to the Lanczos algorithm and modifications of the tridiagonal matrix which is generated by this algorithm and implicitly by CG. This is why we start in Section 2 with the Lanczos algorithm. In Section 3 we discuss the relation with CG and how the Gauss-Radau upper bound is computed. A model problem showing the problems arising with the Gauss-Radau bound in "exact" arithmetic is constructed in Section 4. In Sections 5 to 7 we give an analysis that explains that the problems start when the distance of the smallest Ritz value to the smallest eigenvalue becomes smaller than the distance of μ to the smallest eigenvalue. We also explain why the Gauss-Radau upper bound becomes almost independent of μ. In Section 8 we present an algorithm for improving the upper bounds in previous CG iterations such that the relative accuracy of the upper bounds is guaranteed to be smaller than a prescribed tolerance. Conclusions are given in Section 9.

2 The Lanczos algorithm

Given a starting vector v and a symmetric matrix $A \in \mathbb{R}^{N \times N}$, one can consider a sequence of nested Krylov subspaces

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

The dimension of these subspaces can increase up to an index n called the *grade of v with respect to A*, at which the maximal dimension is attained, and $\mathcal{K}_n(A, v)$ is invariant under multiplication with A.

Algorithm 1 Lanczos algorithm.

> **input** A, v
> $\beta_0 = 0$, $v_0 = 0$
> $v_1 = v/\|v\|$
> **for** $k = 1, \ldots$ **do**
> $w = Av_k - \beta_{k-1}v_{k-1}$
> $\alpha_k = v_k^T w$
> $w = w - \alpha_k v_k$
> $\beta_k = \|w\|$
> $v_{k+1} = w/\beta_k$
> **end for**

Assuming that $k < n$, the Lanczos algorithm (Algorithm 1) computes an orthonormal basis $v_1, \ldots, v_{k+1}$ of the Krylov subspace $\mathcal{K}_{k+1}(A, v)$. The basis vectors v_j satisfy the matrix relation

$$AV_k = V_k T_k + \beta_k v_{k+1} e_k^T$$

where e_k is the last column of the identity matrix of order k, $V_k = [v_1 \cdots v_k]$ and T_k is the $k \times k$ symmetric tridiagonal matrix of the recurrence coefficients computed in Algorithm 1:

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix}.$$

The coefficients β_j being positive, T_k is a so-called Jacobi matrix. If A is positive definite, then T_k is positive definite as well. In the following we will assume for simplicity that the eigenvalues of A are simple and sorted such that

$$\lambda_1 < \lambda_2 < \cdots < \lambda_N.$$

2.1 Approximation of the eigenvalues

The eigenvalues of T_k (Ritz values) are usually used as approximations to the eigenvalues of A. The quality of the approximation can be measured using β_k and the last components of the normalized eigenvectors of T_k. In more detail, consider the spectral decomposition of T_k in the form

$$T_k = S_k \Theta_k S_k^T, \quad \Theta_k = \text{diag}\left(\theta_1^{(k)}, \ldots, \theta_k^{(k)}\right), \quad S_k^T S_k = S_k S_k^T = I_k,$$

where I_k is the identity matrix of order k, and assume that the Ritz values are sorted such that

$$\theta_1^{(k)} < \theta_2^{(k)} < \cdots < \theta_k^{(k)}.$$

Denote $s_{i,j}^{(k)}$ the entries and $s_{:,j}^{(k)}$ the columns of S_k. Then it holds that

$$\min_{i=1,\ldots,N} |\lambda_i - \theta_j^{(k)}| \leq \left\| A\left(V_k s_{:,j}^{(k)}\right) - \theta_j^{(k)}\left(V_k s_{:,j}^{(k)}\right)\right\| = \beta_k |s_{k,j}^{(k)}|, \tag{1}$$

$j = 1, \ldots, k$. Since the Ritz values $\theta_j^{(k)}$ can be seen as Rayleigh quotients, one can improve the bound (1) using the gap theorem; see [14, p. 244] or [15, p. 206]. In particular, let λ_ℓ be an eigenvalue of A closest to $\theta_j^{(k)}$. Then

$$|\lambda_\ell - \theta_j^{(k)}| \leq \frac{\left(\beta_k s_{k,j}^{(k)}\right)^2}{\text{gap}_j^{(k)}}, \qquad \text{gap}_j^{(k)} = \min_{i \neq \ell} |\lambda_i - \theta_j^{(k)}|.$$

In the following we will be interested in the situation when the smallest Ritz value $\theta_1^{(k)}$ closely approximates the smallest eigenvalue of A. If λ_1 is the eigenvalue of A closest to $\theta_1^{(k)} > \lambda_1$, then using the gap theorem and [14, Corollary 11.7.1 on p. 246],

$$\frac{\left(\beta_k s_{k,1}^{(k)}\right)^2}{\lambda_n - \lambda_1} \leq \theta_1^{(k)} - \lambda_1 \leq \frac{\left(\beta_k s_{k,1}^{(k)}\right)^2}{\lambda_2 - \theta_1^{(k)}}, \tag{2}$$

giving the bounds

$$\lambda_2 - \theta_1^{(k)} \leq \frac{\left(\beta_k s_{k,1}^{(k)}\right)^2}{\theta_1^{(k)} - \lambda_1} \leq \lambda_n - \lambda_1. \tag{3}$$

It is known (see, for instance, [16]) that the squares of the last components of the eigenvectors are given by

$$\left(s_{k,j}^{(k)}\right)^2 = \left| \frac{\chi_{1,k-1}(\theta_j^{(k)})}{\chi_{1,k}'(\theta_j^{(k)})} \right|,$$

where $\chi_{1,\ell}$ is the characteristic polynomial of T_ℓ and $\chi_{1,\ell}'$ denotes its derivative, i.e.,

$$\left(s_{k,j}^{(k)}\right)^2 = \frac{\theta_j^{(k)} - \theta_1^{(k-1)}}{\theta_j^{(k)} - \theta_1^{(k)}} \cdots \frac{\theta_j^{(k)} - \theta_{j-1}^{(k-1)}}{\theta_j^{(k)} - \theta_{j-1}^{(k)}} \frac{\theta_j^{(k-1)} - \theta_j^{(k)}}{\theta_{j+1}^{(k)} - \theta_j^{(k)}} \cdots \frac{\theta_{k-1}^{(k-1)} - \theta_j^{[k]}}{\theta_k^{(k)} - \theta_j^{(k)}}.$$

The right-hand side is positive due to the interlacing property of the Ritz values for symmetric tridiagonal matrices. In particular,

$$\left(s_{k,1}^{(k)}\right)^2 = \frac{\theta_1^{(k-1)} - \theta_1^{(k)}}{\theta_2^{(k)} - \theta_1^{(k)}} \cdots \frac{\theta_{k-1}^{(k-1)} - \theta_1^{(k)}}{\theta_k^{(k)} - \theta_1^{(k)}}. \tag{4}$$

When the smallest Ritz value $\theta_1^{(k)}$ converges to λ_1, this last component squared converges to zero; see also (3).

2.2 Modification of the tridiagonal matrix

Given $\mu < \theta_1^{(k)}$, let us consider the problem of finding the coefficient $\alpha_{k+1}^{(\mu)}$ such that the modified matrix

$$
T_{k+1}^{(\mu)} = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \beta_{k-1} & \\ & & \beta_{k-1} & \alpha_k & \beta_k \\ & & & \beta_k & \alpha_{k+1}^{(\mu)} \end{bmatrix}
\tag{5}
$$

has the prescribed μ as an eigenvalue. The connection of this problem to Gauss-Radau quadrature rule will be explained in Section 3.

In [17, pp. 331-334] it has been shown that at iteration $k+1$

$$
\alpha_{k+1}^{(\mu)} = \mu + \zeta_k^{(\mu)}
$$

where $\zeta_k^{(\mu)}$ is the last component of the vector y, solution of the linear system

$$
(T_k - \mu I)y = \beta_k^2 e_k.
\tag{6}
$$

From [10, Section 3.4], the modified coefficients $\alpha_{k+1}^{(\mu)}$ can be computed recursively using

$$
\alpha_{j+1}^{(\mu)} = \mu + \frac{\beta_j^2}{\alpha_j - \alpha_j^{(\mu)}}, \qquad \alpha_1^{(\mu)} = \mu, \qquad j = 1, \dots, k.
\tag{7}
$$

Using the spectral factorization of T_k, we can now prove the following lemma.

Lemma 1 *Let $\mu < \theta_1^{(k)}$. Then it holds that*

$$
\alpha_{k+1}^{(\mu)} = \mu + \sum_{i=1}^{k} \eta_{i,k}^{(\mu)}, \qquad \eta_{i,k}^{(\mu)} \equiv \frac{\left(\beta_k s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}.
\tag{8}
$$

If $\mu < \lambda < \theta_1^{(k)}$, then $\alpha_{k+1}^{(\mu)} < \alpha_{k+1}^{(\lambda)}$. Consequently, if $\mu < \theta_1^{(k+1)}$, then $\alpha_{k+1}^{(\mu)} < \alpha_{k+1}$.

Proof Since $\mu < \theta_1^{(k)}$ the matrix $T_k - \mu I$ in (6) is positive definite and, therefore, nonsingular. Hence,

$$
\zeta_k^{(\mu)} = e_k^T y = \beta_k^2 e_k^T (T_k - \mu I)^{-1} e_k = \sum_{i=1}^{k} \frac{\left(\beta_k s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}
\tag{9}
$$

so that (8) holds. From (8) it is obvious that if $\mu < \lambda < \theta_1^{(k)}$, then $\alpha_{k+1}^{(\mu)} < \alpha_{k+1}^{(\lambda)}$.

Finally, taking $\lambda = \theta_1^{(k+1)} < \theta_1^{(k)}$ (because of the interlacing of the Ritz values) we obtain $\alpha_{k+1}^{(\lambda)} = \alpha_{k+1}$ by construction. □

3 CG and error norm estimation

When solving a linear system $Ax = b$ with a symmetric and positive definite matrix A, the CG method (Algorithm 2) is the method of choice. In exact arithmetic, the CG iterates x_k minimize the A-norm of the error over the manifold $x_0 + \mathcal{K}_k(A, r_0)$,

$$\|x - x_k\|_A = \min_{y \in x_0 + \mathcal{K}_k(A, r_0)} \|x - y\|_A,$$

and the residual vectors $r_k = b - Ax_k$ are proportional to the Lanczos vectors v_j,

$$v_{j+1} = (-1)^j \frac{r_j}{\|r_j\|}, \qquad j = 0, \ldots, k.$$

Thanks to this close relationship between the CG and Lanczos algorithms, it can be shown (see, for instance, [16]) that the recurrence coefficients computed in both algorithms are connected via $\alpha_1 = \gamma_0^{-1}$ and

$$\beta_j = \frac{\sqrt{\delta_j}}{\gamma_{j-1}}, \quad \alpha_{j+1} = \frac{1}{\gamma_j} + \frac{\delta_j}{\gamma_{j-1}}, \quad j = 1, \ldots, k-1. \tag{10}$$

Writing (10) in matrix form, we find out that CG computes implicitly the LDL^T factorization $T_k = L_k D_k L_k^T$, where

$$L_k = \begin{bmatrix} 1 & & & \\ \sqrt{\delta_1} & \ddots & & \\ & \ddots & \ddots & \\ & & \sqrt{\delta_{k-1}} & 1 \end{bmatrix}, \quad D_k = \begin{bmatrix} \gamma_0^{-1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \gamma_{k-1}^{-1} \end{bmatrix}. \tag{11}$$

Algorithm 2 Conjugate gradient algorithm.

---

input A, b, x_0
$r_0 = b - Ax_0$
$p_0 = r_0$
for $k = 1, \ldots$ until convergence **do**

$\quad \gamma_{k-1} = \dfrac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$

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

$\quad r_k = r_{k-1} - \gamma_{k-1} A p_{k-1}$ $\qquad$ cgiter(k-1)

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

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

Hence the matrix T_k is known implicitly in CG.

3.1 Modification of the factorization of T_{k+1}

Similarly as in Section 2.2 we can ask how to modify the Cholesky factorization of T_{k+1}, that is available in CG, such that the resulting matrix $T_{k+1}^{(\mu)}$ given implicitly in factorized form has the prescribed eigenvalue μ. In more detail, we look for a coefficient $\gamma_k^{(\mu)}$ such that

$$T_{k+1}^{(\mu)} = L_{k+1} \begin{bmatrix} D_k & \\ & \left(\gamma_k^{(\mu)}\right)^{-1} \end{bmatrix} L_{k+1}^T.$$

This problem was solved in [10] leading to an updating formula for computing the modified coefficients

$$\gamma_{j+1}^{(\mu)} = \frac{\gamma_j^{(\mu)} - \gamma_j}{\mu(\gamma_j^{(\mu)} - \gamma_j) + \delta_{j+1}}, \quad j = 1, \ldots, k-1, \qquad \gamma_0^{(\mu)} = \frac{1}{\mu}. \qquad (12)$$

Moreover, $\gamma_k^{(\mu)}$ can be obtained directly from the modified coefficient $\alpha_{k+1}^{(\mu)}$,

$$\gamma_k^{(\mu)} = \frac{1}{\alpha_{k+1}^{(\mu)} - \frac{\delta_k}{\gamma_{k-1}}}, \qquad (13)$$

and vice-versa, see [10, p. 173 and 181].

3.2 Quadrature-based bounds in CG

We now briefly summarize the idea of deriving the quadrature-based bounds used in this paper. For a more detailed description, see, e.g., [5–7, 10–13].

Let $A = Q \Lambda Q^T$ be the spectral decomposition of A, with $Q = [q_1, \ldots, q_N]$ orthonormal and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$. As we said above, for simplicity of notation, we assume that the eigenvalues of A are distinct and ordered as $\lambda_1 < \lambda_2 < \cdots < \lambda_N$. Let us define the weights ω_i by

$$\omega_i \equiv \frac{(r_0, q_i)^2}{\|r_0\|^2} \quad \text{so that} \quad \sum_{i=1}^{N} \omega_i = 1 ,$$

and the (nondecreasing) stepwise constant distribution function $\omega(\lambda)$ with a finite number of points of increase $\lambda_1, \lambda_2, \ldots, \lambda_N$,

$$
\omega(\lambda) \equiv
\begin{cases}
0 & \text{for } \lambda < \lambda_1, \\
\sum_{j=1}^{i} \omega_j & \text{for } \lambda_i \leq \lambda < \lambda_{i+1}, \quad 1 \leq i \leq N-1. \\
1 & \text{for } \lambda_N \leq \lambda.
\end{cases}
$$

Having the distribution function $\omega(\lambda)$ and an interval $\langle \eta, \xi \rangle$ such that $\eta < \lambda_1 < \lambda_2 < \cdots < \lambda_N < \xi$, for any continuous function f, one can define the Riemann-Stieltjes integral (see, for instance, [18])

$$
\int_{\eta}^{\xi} f(\lambda) \, d\omega(\lambda) = \sum_{i=1}^{N} \omega_i f(\lambda_i).
$$

For $f(\lambda) = \lambda^{-1}$, we obtain the integral representation of $\|x - x_0\|_A^2$,

$$
\int_{\eta}^{\xi} \lambda^{-1} \, d\omega(\lambda) = \|r_0\|^{-2} \|x - x_0\|_A^2. \tag{14}
$$

Using the optimality of CG it can be shown that CG implicitly determines nodes and weights of the k-point Gauss quadrature approximation to the Riemann-Stieltjes integral (14). The nodes are given by the eigenvalues of T_k, and the weights by the squared first components of the normalized eigenvectors of T_k. The corresponding Gauss quadrature rule can be written in the form

$$
\int_{\eta}^{\xi} \lambda^{-1} \, d\omega(\lambda) = (T_k^{-1})_{1,1} + \frac{\|x - x_k\|_A^2}{\|r_0\|^2}, \tag{15}
$$

where $(T_k^{-1})_{1,1}$ represents the Gauss quadrature approximation, and the reminder is nothing but the scaled and squared A-norm of the kth error, i.e., the quantity of our interest.

To approximate the integral (14), one can also apply a modified quadrature rule. In this paper we consider the Gauss-Radau quadrature rule consisting in prescribing a node $0 < \mu \leq \lambda_1$ and choosing the other nodes and weights to maximize the degree of exactness of the quadrature rule. We can write the corresponding Gauss-Radau quadrature rule in the form

$$
\int_{\eta}^{\xi} \lambda^{-1} \, d\omega(\lambda) = ((T_k^{(\mu)})^{-1})_{1,1} + \mathcal{R}_k^{(\mu)},
$$

where the reminder $\mathcal{R}_k^{(\mu)}$ is negative, and $T_k^{(\mu)}$ is given by (5).

The idea of deriving (basic) quadrature-based bounds in CG is to consider the Gauss quadrature rule (15) at iteration k, and a (eventually modified) quadrature rule at iteration $k + 1$,

$$\frac{\|x - x_0\|_A^2}{\|r_0\|^2} = \left(\widehat{T}_{k+1}^{-1}\right)_{1,1} + \widehat{\mathcal{R}}_{k+1}, \tag{16}$$

where $\widehat{T}_{k+1} = T_{k+1}$ when using the Gauss rule and $\widehat{T}_{k+1} = T_{k+1}^{(\mu)}$ in the case of using the Gauss-Radau rule. From the equations (15) and (16) we get

$$\|x - x_k\|_A^2 = \left[\|r_0\|^2 \left(\left(\widehat{T}_{k+1}^{-1}\right)_{1,1} - \left(T_k^{-1}\right)_{1,1}\right)\right] + \widehat{\mathcal{R}}_{k+1}. \tag{17}$$

The term in square brackets represents either a lower bound on $\|x - x_k\|_A^2$ if $\widehat{T}_{k+1} = T_{k+1}$ (because of the positive reminder), or an upper bound if $\widehat{T}_{k+1} = T_{k+1}^{(\mu)}$ (because of the negative reminder). In both cases, the term in square brackets can easily be evaluated using the available CG related quantities. In particular, the lower bound is given by $\gamma_k \|r_k\|^2$, and the upper bound by $\gamma_k^{(\mu)} \|r_k\|^2$, where $\gamma_k^{(\mu)}$ can be updated using (12).

To summarize results of [5, 7, 12], and [1, 10, 11] related to the Gauss and Gauss-Radau quadrature bounds for the A-norm of the error in CG, it has been shown that

$$\gamma_k \|r_k\|^2 \leq \|x - x_k\|_A^2 < \gamma_k^{(\mu)} \|r_k\|^2 < \left(\frac{\|r_k\|^2}{\mu \|p_k\|^2}\right) \|r_k\|^2 \tag{18}$$

for $k < n - 1$, and μ such that $0 < \mu \leq \lambda_1$. Note that in the special case $k = n - 1$ it holds that $\|x - x_{n-1}\|_A^2 = \gamma_{n-1} \|r_{n-1}\|^2$. If the initial residual r_0 has a nontrivial component in the eigenvector corresponding to λ_1, then λ_1 is an eigenvalue of T_n. If in addition μ is chosen such that $\mu = \lambda_1$, then $\gamma_{n-1} = \gamma_{n-1}^{(\mu)}$ and the second inequality in (18) changes to equality. The last inequality is strict also for $k = n - 1$.

The rightmost bound in (18), that will be called the *simple upper bound* in the following, was derived in [1]. The norm $\|p_k\|$ is not available in CG, but the ratio

$$\phi_k = \frac{\|r_k\|^2}{\|p_k\|^2}$$

can be computed efficiently using

$$\phi_{j+1}^{-1} = 1 + \phi_j^{-1} \delta_{j+1}, \qquad \phi_0 = 1. \tag{19}$$

Note that at an iteration $\ell \leq k$ we can obtain a more accurate bound using

$$\|x - x_\ell\|_A^2 = \sum_{j=\ell}^{k-1} \gamma_j \|r_j\|^2 + \|x - x_k\|_A^2, \tag{20}$$

by applying the basic bounds (18) to the last term in (20); see [1] for details on the construction of more accurate bounds. In practice, however, one runs the CG algorithm, and estimates the error in a backward way, i.e., $k - \ell$ iterations back. The adaptive choice of the delay $k - \ell$ when using the Gauss quadrature lower bound was discussed recently in [19].

In the following we will we concentrate on the analysis of the behavior of the basic Gauss-Radau upper bound

$$\gamma_k^{(\mu)} \|r_k\|^2 \tag{21}$$

in dependence of the choice of μ. As already mentioned, we observed in many examples that in earlier iterations, the bound is approximating the squared A-norm of the error quite well, but in later iterations it becomes worse, it is delayed and almost independent of μ. We observed that this phenomenon is related to the convergence of the smallest Ritz value to the smallest eigenvalue λ_1. In particular, the bound is getting worse if the smallest Ritz value approximates λ_1 better than μ. This often happens during finite precision computations when convergence of CG is delayed because of rounding errors and there are clusters of Ritz values approximating individual eigenvalues of A. Usually, such clusters arise around the largest eigenvalues. At some iteration, each eigenvalue of A can be approximated by a Ritz value, while the A-norm of the error still does not reach the required level of accuracy, and the process will continue and place more and more Ritz values in the clusters. In this situation, it can happen that λ_1 is tightly (that is, to a high relative accuracy) approximated by a Ritz value while the CG process still continues. Note that if A has well separated eigenvalues and we run the experiment in exact arithmetic, then λ_1 is usually tightly approximated by a Ritz value only in the last iterations. The above observation is key for constructing a motivating example, in which we can readily observe the studied phenomenon also in exact arithmetic, and which will motivate our analysis.

4 The model problem and a numerical experiment

In the construction of the motivating example we use results presented in [16, 20–23]. Based on the work by Chris Paige [24], Anne Greenbaum [20] proved that the results of finite precision CG computations can be interpreted (up to some small inaccuracies) as the results of the exact CG algorithm applied to a larger system with the system matrix having many eigenvalues distributed throughout "tiny" intervals around the eigenvalues of the original matrix. The experiments show that "tiny" means of the size comparable to $\mathbf{u}\|A\|$, where $\mathbf{u}$ is the roundoff unit. This result was used in [21] to predict the behavior of finite precision CG. Inspired by [20–22] we will construct a linear system $Ax = b$ with similar properties as the one suggested by Greenbaum [20]. However, we want to emphasize and visualize some phenomenons concerning the behavior of the basic Gauss-Radau upper bound (21). Therefore, we choose the size of the intervals around the original eigenvalues larger than $\mathbf{u}\|A\|$.

We start with the test problem $\Lambda y = w$ from [23], where $w = m^{-1/2}(1, \ldots, 1)^T$ and $\Lambda = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$,

$$\hat{\lambda}_i = \hat{\lambda}_1 + \frac{i-1}{m-1}(\hat{\lambda}_m - \hat{\lambda}_1)\rho^{m-i}, \quad i = 2, \ldots, m. \tag{22}$$

The diagonal matrix Λ and the vector w determine the stepwise distribution function $\omega(\lambda)$ with points of increase $\hat{\lambda}_i$ and the individual jumps (weights) $\omega_j = m^{-1}$,

$$\omega(\lambda) \equiv \begin{cases} 0 & \text{for } \lambda < \hat{\lambda}_1, \\ \sum_{j=1}^{i} \omega_j & \text{for } \hat{\lambda}_i \le \lambda < \hat{\lambda}_{i+1}, \quad 1 \le i \le m-1, \\ 1 & \text{for } \hat{\lambda}_m \le \lambda. \end{cases} \tag{23}$$

We construct a blurred distribution function $\widetilde{\omega}(\lambda)$ having clusters of points of increase around the original eigenvalues $\hat{\lambda}_i$. We consider each cluster to have the same radius δ, and let the number c_i of points in the ith cluster grow linearly from 1 to p,

$$c_i = \mathrm{round}\left(\frac{p-1}{m-1}i + \frac{m-p}{m-1}\right), \quad i = 1, \ldots, m.$$

The blurred eigenvalues

$$\widetilde{\lambda}_j^{(i)}, \quad j = 1, \ldots, c_i,$$

are uniformly distributed in $[\hat{\lambda}_i - \delta, \hat{\lambda}_i + \delta]$, with the corresponding weights given by

$$\widetilde{\omega}_j^{(i)} = \frac{\omega_i}{c_i} \quad j = 1, \ldots, c_i,$$

i.e., the weights that correspond to the ith cluster are equal, and their sum is ω_i. Having defined the blurred distribution function $\widetilde{\omega}(\lambda)$ we can construct the corresponding Jacobi matrix $T \in \mathbb{R}^{N \times N}$ in a numerically stable way using the Gragg and Harrod rkpw algorithm [25]. Note that the mapping from the nodes and weights of the computed quadrature to the recurrence coefficients is generally well-conditioned [26, p. 59]. To construct the above-mentioned Jacobi matrix T we used Matlab's vpa arithmetic with 128 digits. Finally, we define the *double precision data A and b* that will be used for experimenting as

$$A = \mathrm{double}(T), \quad b = e_1, \tag{24}$$

where $e_1 \in \mathbb{R}^N$ is the first column of the identity matrix. We decided to use double precision input data since we can easily compare results of our computations performed in Matlab's vpa arithmetic with results obtained using double precision arithmetic for the same input data.

In our experiment we choose $m = 12, \hat{\lambda}_1 = 10^{-6}, \hat{\lambda}_m = 1, \rho = 0.8, \delta = 10^{-10}$, and $p = 4$, resulting in $N = 30$. Let us run the "exact" CGQ algorithm of [10] on the model problem (24) constructed above, where exact arithmetic is simulated using Matlab's variable precision with `digits=128`. Let λ_1 be the exact smallest eigenvalue of A. We use four different values of μ for the computation of the Gauss-Radau upper bound

(21): $\mu_3 = (1-10^{-3})\lambda_1$, $\mu_8 = (1-10^{-8})\lambda_1$, μ_{16} which denotes the double precision number closest to λ_1 such that $\mu_{16} \le \lambda_1$, and $\mu_{50} = (1-10^{-50})\lambda_1$ which is almost like the exact value. Note that $\gamma_k^{(\mu)}$ is updated using (12).

Figure 1 shows the A-norm of the error $\|x - x_{k-1}\|_A$ (solid curve), the upper bounds for the considered values of μ_i, $i = 3, 8, 16, 50$ (dotted solid curves), and the rightmost bound in (18) (the simple upper bound) for μ_{50} (dashed curve). The dots represent the values $\theta_1^{(k)} - \lambda_1$, i.e., the distances of the smallest Ritz values $\theta_1^{(k)}$ to λ_1. The horizontal dotted lines correspond to the values of $\lambda_1 - \mu_i$, $i = 3, 8, 16$.

The Gauss-Radau upper bounds in Fig. 1 first overestimate, and then closely approximate $\|x-x_{k-1}\|_A$ (starting from iteration 5). However, at some point, the Gauss-Radau upper bounds start to differ significantly from $\|x-x_{k-1}\|_A$ and represent worse approximations, except for μ_{50}. We observe that for a given μ_i, $i = 3, 8, 16$, the upper bounds are delayed when the distance of $\theta_1^{(k)}$ to λ_1 becomes smaller than the distance of μ_i to λ_1, i.e., when

$$\theta_1^{(k)} - \lambda_1 < \lambda_1 - \mu_i. \tag{25}$$

If (25) holds, then the smallest Ritz value $\theta_1^{(k)}$ is a better approximation to λ_1 than μ_i. This moment is emphasized using vertical dashed lines that connect the value $\theta_1^{(k)} - \lambda_1$ with $\|x-x_{k-1}\|_A$ in the first iteration k such that (25) holds. Moreover, below a certain level, the upper bounds become almost independent of μ_i, $i = 3, 8, 16$, and visually coincide with the simple upper bound. The closer is μ to λ_1, the later this phenomenon occurs.

Depending on the validity of (25), we distinguish between *phase 1* and *phase 2* of convergence. If the inequality (25) does not hold, i.e., if μ is a better approximation to λ_1 than the smallest Ritz value, then we say we are in phase 1. If (25) holds, then the smallest Ritz value is closer to λ_1 than μ and we are in phase 2.
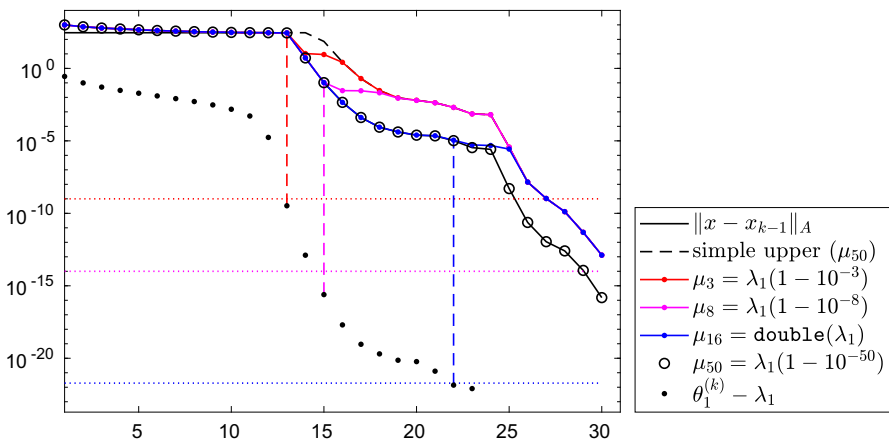

Fig. 1 $\|x - x_{k-1}\|_A$, upper bounds and the distance of $\theta_1^{(k)}$ to λ_1, digits=128

Obviously, the beginning of phase 2 depends on the choice of μ and on the convergence of the smallest Ritz value to the smallest eigenvalue. Note that for $\mu = \mu_{50}$ we are always in phase 1 before we stop the iterations.

In the given experiment as well as in many practical problems, the delay of the upper bounds is not large (just a few iterations), and the bounds can still provide a useful information for stopping the algorithm. However, we have also encountered examples where the delay of the Gauss-Radau upper bound was about 200 iterations; see, e.g., [1, Fig. 10] or [19, Fig. 2] concerning the matrix s3dkt3m2. Hence, we believe that this phenomenon deserves attention and explanation.

5 Analysis

The upper bounds are computed from the modified tridiagonal matrices (5) discussed in Section 2.2, that differ only in one coefficient at the position $(k+1, k+1)$. Therefore, the first step of the analysis is to understand how the choice of μ and the validity of the condition (25) influences the value of the modified coefficient

$$\alpha_{k+1}^{(\mu)} = \mu + \sum_{i=1}^{k} \eta_{i,k}^{(\mu)}, \qquad \eta_{i,k}^{(\mu)} = \frac{\left(\beta_k s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}; \tag{26}$$

see (8). We will compare its value to a modified coefficient for which phase 2 does not occur; see Fig. 1 for μ_{50}.

Based on that understanding we will then address further important questions. First, our aim is to explain the behavior of the basic Gauss-Radau upper bound (21) in phase 2, in particular, its closeness to the simple upper bound (18). Second, for practical reasons, without knowing λ_1, we would like to be able to detect phase 2, i.e., the first iteration k for which the inequality (25) starts to hold. Finally, we address the problem of how to improve the accuracy of the basic Gauss-Radau upper bound (21) in phase 2.

We first analyze the relation between two modified coefficients $\alpha_{k+1}^{(\mu)}$ and $\alpha_{k+1}^{(\lambda)}$ where $0 < \mu < \lambda < \theta_1^{(k)}$.

Lemma 2 *Let* $0 < \mu < \lambda < \theta_1^{(k)}$. *Then*

$$\frac{\eta_{i,k}^{(\lambda)} - \eta_{i,k}^{(\mu)}}{\eta_{i,k}^{(\mu)}} = \frac{\lambda - \mu}{\theta_i^{(k)} - \lambda} \tag{27}$$

and

$$\alpha_{k+1}^{(\lambda)} - \alpha_{k+1}^{(\mu)} = \left(\frac{\lambda - \mu}{\theta_1^{(k)} - \mu}\right) \eta_{1,k}^{(\lambda)} + (\lambda - \mu) E_k^{(\lambda,\mu)} \tag{28}$$

where

$$E_k^{(\lambda,\mu)} \equiv 1 + \sum_{i=2}^{k} \frac{\eta_{i,k}^{(\lambda)}}{\theta_i^{(k)} - \mu} \tag{29}$$

satisfies $E_k^{(\lambda,\mu)} = E_k^{(\mu,\lambda)}$.

Proof From the definition of $\eta_{i,k}^{(\mu)}$ and $\eta_{i,k}^{(\lambda)}$, it follows immediately

$$\frac{\eta_{i,k}^{(\lambda)}}{\theta_i^{(k)} - \mu} = \frac{\eta_{i,k}^{(\mu)}}{\theta_i^{(k)} - \lambda},$$

which implies $E_k^{(\lambda,\mu)} = E_k^{(\mu,\lambda)}$ and (27).

Note that $0 < \eta_{i,k}^{(\mu)} < \eta_{i,k}^{(\lambda)}$. Using (27), the difference of the coefficients α's is

$$\alpha_{k+1}^{(\lambda)} - \alpha_{k+1}^{(\mu)} = (\lambda - \mu) + \sum_{i=1}^{k} \left(\eta_{i,k}^{(\lambda)} - \eta_{i,k}^{(\mu)} \right)$$

$$= (\lambda - \mu) + (\lambda - \mu) \sum_{i=1}^{k} \frac{\eta_{i,k}^{(\mu)}}{\theta_i^{(k)} - \lambda}$$

$$= (\lambda - \mu) \frac{\eta_{1,k}^{(\lambda)}}{\theta_1^{(k)} - \mu} + (\lambda - \mu) \left(1 + \sum_{i=2}^{k} \frac{\eta_{i,k}^{(\mu)}}{\theta_i^{(k)} - \lambda} \right)$$

which implies (28). □

5.1 Assumptions

Let us describe in more detail the situation we are interested in. In the analysis that follows we will assume implicitly the following.

1. λ_1 is well separated from λ_2 so that we can use the gap theorem mentioned in Section 2.1, in particular relation (3) bounding $\eta_{1,k}^{(\lambda_1)}$.
2. μ is a tight underestimate to λ_1 such that

$$\lambda_1 - \mu \ll \lambda_2 - \lambda_1. \tag{30}$$

3. The smallest Ritz value $\theta_1^{(k)}$ converges to λ_1 with increasing k so that there is an iteration index k from which

$$\theta_1^{(k)} - \lambda_1 \ll \lambda_1 - \mu.$$

Let us briefly comment on these assumptions. The assumption that λ_1 is well separated from λ_2 is used later to prove that $\eta_{1,k}^{(\lambda_1)}$ is bounded away from zero; see (33).

If there is a cluster of eigenvalues around λ_1, one can still observe the discussed phenomenon of loss of accuracy of the upper bound, but a theoretical analysis would be much more complicated. Note that the first assumption is also often satisfied for a system matrix $\hat{A}$ that models finite precision CG behavior, if the original matrix A has well separated eigenvalues λ_1 and λ_2. Using results of Greenbaum [20] we know that $\hat{A}$ can have many eigenvalues distributed throughout tiny intervals around the eigenvalues of A. We have constructed the model matrix $\hat{A}$ in many numerical experiments, using the procedure suggested in [20]. We found out that the constructed $\hat{A}$ has usually clusters of eigenvalues around the larger eigenvalues of A while a smaller eigenvalue of A is usually approximated by just one eigenvalue of $\hat{A}$. Therefore, the analysis presented below can then be applied to the matrix $\hat{A}$ that models the finite precision CG behavior.

If μ is not a tight underestimate, then the Gauss-Radau upper bound is usually not a very good approximation of the A-norm of the error. Then the condition (25) can hold from the beginning and phase 1 need not happen.

Finally, in theory, the smallest Ritz value need not converge to λ_1 until the last iteration [27]. But, in that case, there won't be any problem for the Gauss-Radau upper bound. However, in practical computations, we very often observe the convergence of $\theta_1^{(k)}$ to λ_1. In particular, in cases of matrices $\hat{A}$ with clustered eigenvalues that model finite precision behavior of CG, $\theta_1^{(k)}$ approximates λ_1 to a high relative accuracy usually earlier before the A-norm of the error reaches the ultimate level of accuracy.

5.2 The modified coefficient $\alpha_{k+1}^{(\mu)}$

Below we would like to compare $\alpha_{k+1}^{(\lambda_1)}$ for which phase 2 does not occur with $\alpha_{k+1}^{(\mu)}$ for which phase 2 occurs; see Fig. 1. Using (27) and (30), we are able to compare the individual η-terms. In particular, for $i > 1$ we get

$$\frac{\eta_{i,k}^{(\lambda_1)} - \eta_{i,k}^{(\mu)}}{\eta_{i,k}^{(\mu)}} = \frac{\lambda_1 - \mu}{\theta_i^{(k)} - \lambda_1} < \frac{\lambda_1 - \mu}{\lambda_2 - \lambda_1} \ll 1,$$

where we have used $\theta_i^{(k)} > \lambda_2$ for $i > 1$. Therefore,

$$\eta_{i,k}^{(\lambda_1)} \approx \eta_{i,k}^{(\mu)} \quad \text{for} \quad i > 1.$$

Hence, $\alpha_{k+1}^{(\mu)}$ can significantly differ from $\alpha_{k+1}^{(\lambda_1)}$ only in the first term of the sum in (26) for which

$$\frac{\eta_{1,k}^{(\lambda_1)} - \eta_{1,k}^{(\mu)}}{\eta_{1,k}^{(\mu)}} = \frac{\lambda_1 - \mu}{\theta_1^{(k)} - \lambda_1}. \tag{31}$$

If $\theta_1^{(k)}$ is a better approximation to λ_1 than μ in the sense of (25), then (31) shows that $\eta_{1,k}^{(\lambda_1)}$ can be much larger than $\eta_{1,k}^{(\mu)}$. As a consequence, $\alpha_{k+1}^{(\lambda_1)}$ can differ significantly from $\alpha_{k+1}^{(\mu)}$. On the other hand, if μ is chosen such that

$$\lambda_1 - \mu \ll \theta_1^{(k)} - \lambda_1,$$

for all k we are interested in, then phase 2 will not occur, and

$$\alpha_{k+1}^{(\lambda_1)} - \alpha_{k+1}^{(\mu)} = (\lambda_1 - \mu) + \sum_{i=1}^{k} \left(\eta_{i,k}^{(\lambda_1)} - \eta_{i,k}^{(\mu)} \right) \approx 0,$$

since μ is assumed to be a tight approximation to λ_1 and $\eta_{i,k}^{(\lambda_1)} \approx \eta_{i,k}^{(\mu)}$ for all i.

In the following we discuss phase 1 and phase 2 in more detail.

In phase 1,

$$\lambda_1 - \mu < \theta_1^{(k)} - \lambda_1,$$

and, therefore, all components $\eta_{i,k}^{(\mu)}$ (including $\eta_{1,k}^{(\mu)}$) are not sensitive to small changes of μ; see (27). In other words, the coefficients $\alpha_{k+1}^{(\mu)}$ are approximately the same for various choices of μ.

Let us denote

$$h_k = \frac{\lambda_1 - \mu}{\theta_1^{(k)} - \lambda_1} < 1.$$

In fact, we can write $\theta_1^{(k)} - \mu = \theta_1^{(k)} - \lambda_1 + \lambda_1 - \mu$ and use the Taylor expansion of $1/(1 + h_k)$. It yields

$$\frac{1}{\theta_1^{(k)} - \mu} = \frac{1}{\theta_1^{(k)} - \lambda_1} \left(\frac{1}{h_k + 1} \right) = \frac{1}{\theta_1^{(k)} - \lambda_1} \left[1 - h_k + h_k^2 - h_k^3 + \cdots \right].$$

Obviously, h_k is an increasing function of the iteration number k; the numerator is constant while the denominator is decreasing in absolute value. The size of h_k depends also on how well μ approximates λ_1. If μ is a tight approximation to λ_1, then, at the beginning of the CG iterations, the denominator of h_k can be large compared to the numerator, h_k is small and the right-hand side of $1/(\theta_1^{(k)} - \mu)$ is almost given by $1/(\theta_1^{(k)} - \lambda_1)$, independent of μ. We observed that the first term of the sum of the $\eta_{i,k}^{(\mu)}$ is then usually the largest one.

Let us now discuss phase 2. First recall that for any $0 < \mu < \lambda_1$ it holds that

$$\alpha_{k+1}^{(\mu)} < \alpha_{k+1}^{(\lambda_1)} \quad \text{and} \quad \eta_{1,k}^{(\mu)} < \eta_{1,k}^{(\lambda_1)}. \tag{32}$$

As before, suppose that λ_1 is well separated from λ_2 and that (30) holds. Phase 2 begins when $\theta_1^{(k)}$ is a better approximation to λ_1 than μ, i.e., when (25) holds. Since $\theta_1^{(k)}$ is a tight approximation to λ_1 in phase 2, (3) and (25) imply that

$$\eta_{1,k}^{(\lambda_1)} \geq \lambda_2 - \theta_1^{(k)} = \lambda_2 - \lambda_1 + \lambda_1 - \theta_1^{(k)} > (\lambda_2 - \lambda_1) - (\lambda_1 - \mu). \qquad (33)$$

Therefore, using (30), $\eta_{1,k}^{(\lambda_1)}$ is bounded away from zero. On the other hand, (3) also implies that

$$\eta_{1,k}^{(\mu)} = \frac{\theta_1^{(k)} - \lambda_1}{\theta_1^{(k)} - \mu} \eta_{1,k}^{(\lambda_1)} \leq \frac{\theta_1^{(k)} - \lambda_1}{\theta_1^{(k)} - \mu} (\lambda_n - \lambda_1)$$

and as $\theta_1^{(k)}$ converges to λ_1, $\eta_{1,k}^{(\mu)}$ goes to zero. Therefore,

$$\alpha_{k+1}^{(\mu)} \approx \mu + \sum_{i=2}^{k} \eta_{i,k}^{(\mu)},$$

and the sum on the right-hand side is almost independent of μ. Note that having two values $0 < \mu < \lambda < \lambda_1$ such that

$$\theta_1^{(k)} - \lambda_1 < \lambda_1 - \lambda \quad \text{and} \quad \lambda - \mu \ll \lambda_2 - \lambda_1, \qquad (34)$$

then one can expect that

$$\alpha_{k+1}^{(\mu)} \approx \alpha_{k+1}^{(\lambda)} \qquad (35)$$

because $\eta_{1,k}^{(\mu)}$ and $\eta_{1,k}^{(\lambda)}$ converge to zero and $\eta_{i,k}^{(\mu)} \approx \eta_{i,k}^{(\lambda)}$ for $i > 1$ due to

$$\frac{\eta_{i,k}^{(\lambda)} - \eta_{i,k}^{(\mu)}}{\eta_{i,k}^{(\mu)}} = \frac{\lambda - \mu}{\theta_i^{(k)} - \lambda} < \frac{\lambda - \mu}{\lambda_2 - \lambda_1} \ll 1,$$

where we have used (27) and the assumption (34). Therefore, $\alpha_{k+1}^{(\mu)}$ is relatively insensitive to small changes of μ and the same is true for the upper bound (21).

5.3 The coefficient α_{k+1}

The coefficient α_{k+1} can also be written as

$$\alpha_{k+1} = \alpha_{k+1}^{(\mu)} \quad \text{for} \quad \mu = \theta_1^{(k+1)},$$

and the results of Lemmas 1 and 2 are still valid, even though, in practice, μ must be smaller than λ_1. Using (28) we can express the differences between the coefficients, it holds that

$$\alpha_{k+1} - \alpha_{k+1}^{(\lambda_1)} = \eta_{1,k}^{(\lambda_1)} \frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \theta_1^{(k+1)}} + \left(\theta_1^{(k+1)} - \lambda_1\right) E_k^{(\theta_1^{(k+1)}, \lambda_1)}. \tag{36}$$

If the smallest Ritz value $\theta_1^{(k+1)}$ is close to λ_1, then the second term of the right-hand side in (36) will be negligible in comparison to the first one, since

$$E_k^{(\theta_1^{(k+1)}, \lambda_1)} = \mathcal{O}(1),$$

see (29), and since $\eta_{1,k}^{(\lambda_1)}$ is bounded away from zero; see (33). Therefore, one can expect that

$$\alpha_{k+1} - \alpha_{k+1}^{(\lambda_1)} \approx \eta_{1,k}^{(\lambda_1)} \frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \theta_1^{(k+1)}}. \tag{37}$$

The size of the term on the right-hand side is related to the speed of convergence of the smallest Ritz value $\theta_1^{(k)}$ to λ_1. Denoting

$$\frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \lambda_1} = \rho_k < 1,$$

we obtain

$$\frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \theta_1^{(k+1)}} = \frac{\frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \lambda_1}}{1 - \frac{\theta_1^{(k+1)} - \lambda_1}{\theta_1^{(k)} - \lambda_1}} = \frac{\rho_k}{1 - \rho_k}.$$

For example, if the convergence of $\theta_1^{(k)}$ to λ_1 is superlinear, i.e., if $\rho_k \to 0$, then α_{k+1} and $\alpha_{k+1}^{(\lambda_1)}$ are close.

5.4 Numerical experiments

Let us demonstrate numerically the theoretical results described in previous sections using our model problem. To compute the following results, we, again, use Matlab's vpa arithmetic with 128 decimal digits.

We first consider $\mu = \mu_3 = (1 - 10^{-3})\lambda_1$ for which we have $\lambda_1 - \mu = 10^{-9}$. The switch from phase 1 to phase 2 occurs at iteration 13. Figure 2 displays the first term $\eta_{1,k}^{(\mu)}$ and the maximum term $\eta_{i,k}^{(\mu)}$ as well as the sum $\zeta_k^{(\mu)}$ defined by (9), see Lemma 1, as a function of the iteration number k. In phase 1 the first term $\eta_{1,k}^{(\mu)}$ is the largest one. As predicted, after the start of phase 2, the first term is decreasing quite fast.
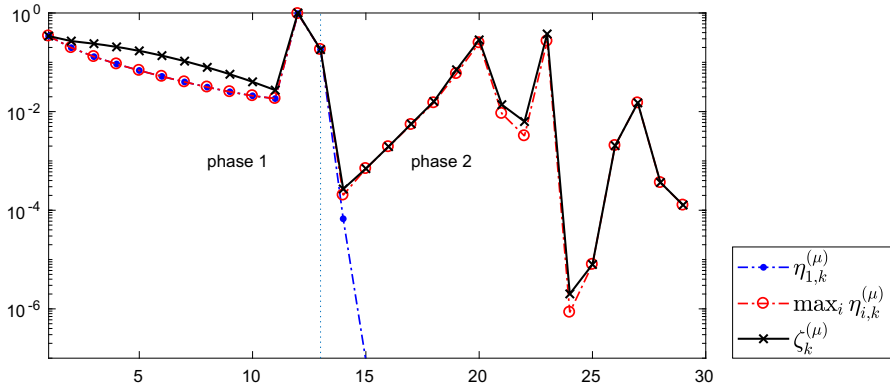
Fig. 2 First term $\eta_{1,k}^{(\mu)}$, maximum term $\eta_{i,k}^{(\mu)}$, and the sum $\zeta_k^{(\mu)}$ for $\mu = \mu_3$

Let us now use $\mu = \mu_8 = (1 - 10^{-8})\lambda_1$ for which we have $\lambda_1 - \mu = 10^{-14}$. The switch from phase 1 to phase 2 occurs at iteration 15; see Fig. 3. The conclusions are the same as for μ_3.

The behavior of the first term is completely different for $\mu = (1 - 10^{-50})\lambda_1$ which almost corresponds to using the exact smallest eigenvalue λ_1.

The maximum term of the sum is then almost always the first one; see Fig. 4. Remember that, for this value of μ, we are always in phase 1.

Finally, in Fig. 5 we present a comparison of the sums $\zeta_k^{(\mu)}$ for μ_3, μ_8, and μ_{50}. We observe that from the beginning up to iteration 12, all sums visually coincide. Starting from iteration 13 we enter phase 2 for $\mu = \mu_3$ and the sum $\zeta_k^{(\mu_3)}$ starts to differ significantly from the other sums, in particular from the "reference" term $\zeta_k^{(\mu_{50})}$. Similarly, for $k = 15$ we enter phase 2 for $\mu = \mu_8$ and $\zeta_k^{(\mu_8)}$ and $\zeta_k^{(\mu_{50})}$ start to differ. We can also observe that $\zeta_k^{(\mu_3)}$ and $\zeta_k^{(\mu_8)}$ significantly differ only in iterations 13, 14,
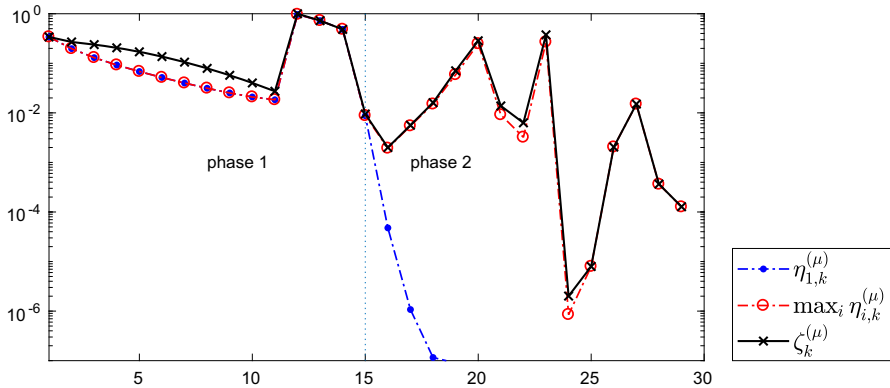

Fig. 3 First term $\eta_{1,k}^{(\mu)}$, maximum term $\eta_{i,k}^{(\mu)}$, and the sum $\zeta_k^{(\mu)}$ for $\mu = \mu_8$
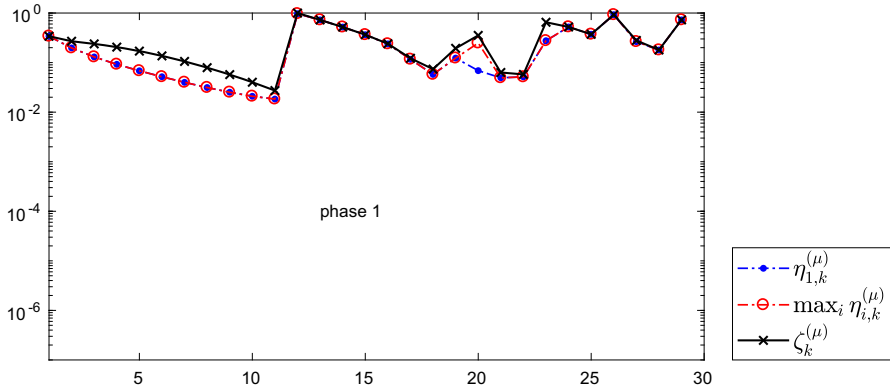
Fig. 4 First term $\eta_{1,k}^{(\mu)}$, maximum term $\eta_{i,k}^{(\mu)}$, and the sum $\zeta_k^{(\mu)}$ for $\mu = \mu_{50}$

and 15, i.e., when we are in phase 2 for $\mu = \mu_3$ but in phase 1 for $\mu = \mu_8$. In all other iterations, $\zeta_k^{(\mu_3)}$ and $\zeta_k^{(\mu_8)}$ visually coincide.

In Fig. 6 we plot the coefficients $\alpha_k^{(\mu_3)}$, $\alpha_k^{(\mu_8)}$, $\alpha_k^{(\lambda_1)}$ and α_k, so that we can compare the observed behavior with the predicted one. Phase 2 starts for μ_3 at iteration 13, and for μ_8 at iteration 15; see also Fig. 1. For $k \leq 13$ we observe that

$$\alpha_k^{(\mu_3)} \approx \alpha_k^{(\mu_8)} \approx \alpha_k^{(\lambda_1)}$$

as explained in Section 5.2 and α_k is larger. For $k \geq 16$, the first terms $\eta_{1,k-1}^{(\mu_3)}$ and $\eta_{1,k-1}^{(\mu_8)}$ are close to zero, and, as explained in Section 5.2,

$$\alpha_k^{(\mu_3)} \approx \alpha_k^{(\mu_8)}.$$
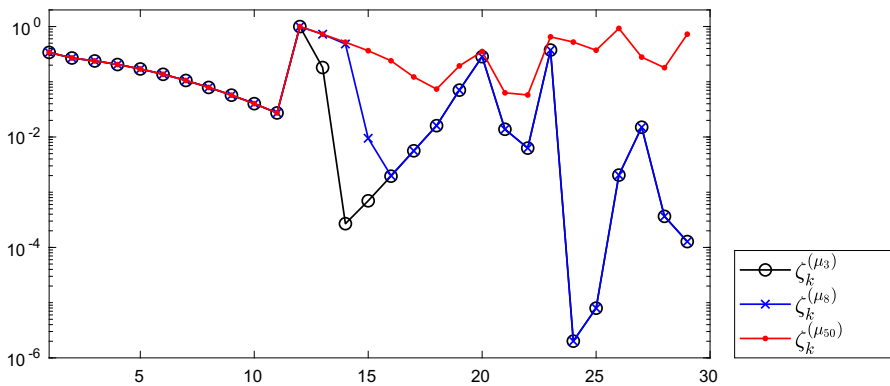
Fig. 5 Comparison of the sums $\zeta_k^{(\mu_3)}$, $\zeta_k^{(\mu_8)}$, and $\zeta_k^{(\mu_{50})}$
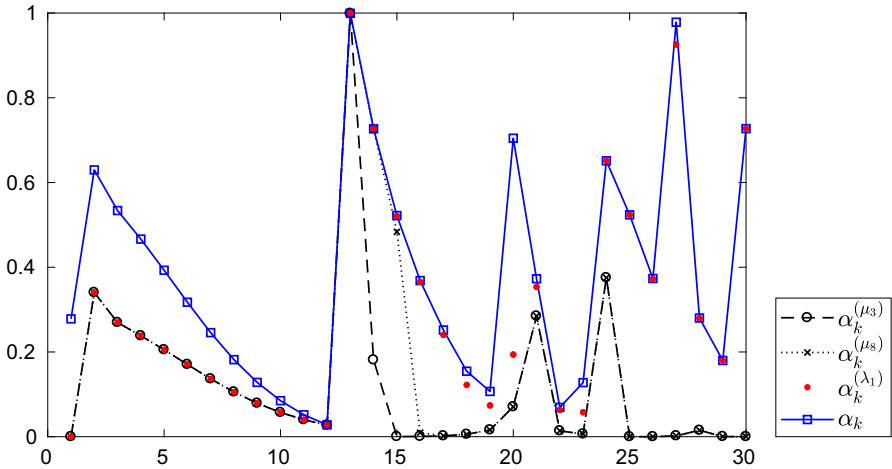
Fig. 6 $\alpha_k^{(\mu_3)}, \alpha_k^{(\mu_8)}, \alpha_k^{(\lambda_1)}$, and α_k

For $k = 14$ and $k = 15$, $\alpha_k^{(\mu_3)}$ and $\alpha_k^{(\mu_8)}$ can differ significantly because $\alpha_k^{(\mu_3)}$ is already in phase 2 while $\alpha_k^{(\mu_8)}$ is still in phase 1.

We can also observe that α_k can be very close to $\alpha_k^{(\lambda_1)}$ when the smallest Ritz value $\theta_1^{(k)}$ is a tight approximation to λ_1, i.e., in later iterations. We know that the closeness of α_k to $\alpha_k^{(\lambda_1)}$ depends on the speed of convergence of the smallest Ritz value to λ_1; see (37) and the corresponding discussion.

6 The Gauss-Radau bound in phase 2

Our aim in this section is to investigate the relation between the basic Gauss-Radau upper bound (21) and the simple upper bound; see (18). Recall the notation

$$\phi_k = \frac{\|r_k\|^2}{\|p_k\|^2};$$

see (19). In particular, we would like to explain why the two bounds almost coincide in phase 2. Note that using (13) we obtain

$$\alpha_{k+1}^{(\mu)} = \left(\gamma_k^{(\mu)}\right)^{-1} + \frac{\delta_k}{\gamma_{k-1}} \tag{38}$$

and from (8) it follows

$$\alpha_{k+1}^{(\mu)} = \mu + \beta_k^2 e_k^T \left(T_k - \mu I\right)^{-1} e_k, \qquad \beta_k^2 = \frac{1}{\gamma_{k-1}} \frac{\delta_k}{\gamma_{k-1}}.$$

Therefore,

$$\left(\gamma_k^{(\mu)}\right)^{-1} = \mu + \beta_k^2 \left(e_k^T \ (T_k - \mu I)^{-1} \ e_k - \gamma_{k-1}\right). \tag{39}$$

In the following lemma we give another expression for $e_k^T \ (T_k - \mu I)^{-1} \ e_k$.

Lemma 3 *Let* $0 < \mu < \theta_1^{(k)}$. *Then it holds that*

$$e_k^T \ (T_k - \mu I)^{-1} \ e_k = \gamma_{k-1} + \mu \frac{\gamma_{k-1}^2}{\phi_{k-1}} + \sum_{i=1}^k \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{\left(s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}. \tag{40}$$

Proof Since $\left\|\mu T_k^{-1}\right\| < 1$, we obtain using a Neumann series

$$(T_k - \mu I)^{-1} = \left(I - \mu T_k^{-1}\right)^{-1} T_k^{-1} = \left(\sum_{j=0}^\infty \mu^j T_k^{-j}\right) T_k^{-1}$$

so that

$$e_k^T \ (T_k - \mu I)^{-1} \ e_k = e_k^T T_k^{-1} e_k + \mu e_k^T T_k^{-2} e_k + \sum_{j=2}^\infty \mu^j e_k^T T_k^{-(j+1)} e_k.$$

We now express the terms on the right-hand side using the CG coefficients and the quantities from the spectral factorization of T_k. Using $T_k = L_k D_k L_k^T$ we obtain $e_k^T T_k^{-1} e_k = \gamma_{k-1}$. After some algebraic manipulation, see, e.g., [28, p. 1369] we get

$$T_k^{-1} e_k = \gamma_{k-1} \|r_{k-1}\| \begin{bmatrix} \frac{(-1)^{k-1}}{\|r_0\|} \\ \vdots \\ \frac{1}{\|r_{k-1}\|} \end{bmatrix}$$

so that

$$e_k^T T_k^{-2} e_k = e_k^T T_k^{-1} T_k^{-1} e_k = \gamma_{k-1}^2 \sum_{i=0}^{k-1} \frac{\|r_{k-1}\|^2}{\|r_i\|^2} = \gamma_{k-1}^2 \frac{\|p_{k-1}\|^2}{\|r_{k-1}\|^2} = \frac{\gamma_{k-1}^2}{\phi_{k-1}}.$$

Finally,

$$e_k^T \left(\sum_{j=2}^\infty \mu^j T_k^{-(j+1)}\right) e_k = e_k^T S_k \left(\sum_{j=2}^\infty \mu^j \Theta_k^{-(j+1)}\right) S_k^T e_k$$

where the diagonal entries of the diagonal matrix

$$\sum_{j=2}^{\infty} \mu^j \Theta_k^{-(j+1)}$$

have the form

$$\frac{1}{\theta_i^{(k)}} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \sum_{j=0}^{\infty} \left(\frac{\mu}{\theta_i^{(k)}}\right)^j = \frac{1}{\theta_i^{(k)}} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{1}{1 - \frac{\mu}{\theta_i^{(k)}}} = \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{1}{\theta_i^{(k)} - \mu}.$$

Hence,

$$e_k^T \left(\sum_{j=2}^{\infty} \mu^j T_k^{-(j+1)}\right) e_k = \sum_{i=1}^{k} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{\left(s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}.$$

$\square$

Based on the previous lemma we can now express the coefficient $\gamma_k^{(\mu)}$.

Theorem 1 *Let $0 < \mu < \theta_1^{(k)}$. Then it holds that*

$$\left(\gamma_k^{(\mu)}\right)^{-1} = \frac{\mu}{\phi_k} + \sum_{i=1}^{k} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \eta_{i,k}^{(\mu)}. \tag{41}$$

Proof We start with (39). Using the previous lemma

$$\left(\gamma_k^{(\mu)}\right)^{-1} = \mu + \mu\beta_k^2 \gamma_{k-1}^2 \phi_{k-1}^{-1} + \beta_k^2 e_k^T \left(\sum_{j=2}^{\infty} \mu^j T_k^{-(j+1)}\right) e_k$$

$$= \mu\left(1 + \delta_k \phi_{k-1}^{-1}\right) + \beta_k^2 \sum_{i=1}^{k} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{\left(s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu}$$

$$= \mu\phi_k^{-1} + \sum_{i=1}^{k} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \frac{\left(\beta_k s_{k,i}^{(k)}\right)^2}{\theta_i^{(k)} - \mu},$$

where we have used relation (19). $\square$

Obviously, using (41), the basic Gauss-Radau upper bound (21) and the simple upper bound in (18) are close to each other if and only if

$$\sum_{i=1}^{k} \left(\frac{\mu}{\theta_i^{(k)}}\right)^2 \eta_{i,k}^{(\mu)} \ll \frac{\mu}{\phi_k}, \tag{42}$$

which can also be written as

$$\left(\frac{\mu}{\theta_1^{(k)}}\right)^2 \frac{\eta_{1,k}^{(\mu)}}{\mu} + \sum_{i=2}^{k} \left(\frac{\beta_k s_{k,i}^{(k)}}{\theta_i^{(k)}}\right)^2 \frac{\mu}{\theta_i^{(k)} - \mu} \ll \phi_k^{-1}. \tag{43}$$

Under the assumptions formulated in Section 5.1, in particular that λ_1 is well separated from λ_2, and that μ is a tight underestimate to λ_1 in the sense of (30), the sum of terms on the left-hand side of (43) can be replaced by its tight upper bound

$$\left(\frac{\lambda_1}{\theta_1^{(k)}}\right)^2 \frac{\eta_{1,k}^{(\mu)}}{\mu} + \sum_{i=2}^{k} \left(\frac{\beta_k s_{k,i}^{(k)}}{\theta_i^{(k)}}\right)^2 \frac{\lambda_1}{\theta_i^{(k)} - \lambda_1} \tag{44}$$

which simplifies the explanation of the dependence of the sum in (43) on μ.

The second term in (44) is independent of μ and its size depends only on the behavior of the underlying Lanczos process. Here

$$\left(\frac{\beta_k s_{k,i}^{(k)}}{\theta_i^{(k)}}\right)^2 = \frac{\left\| A\left(V_k s_{:,i}^{(k)}\right) - \theta_i^{(k)} \left(V_k s_{:,i}^{(k)}\right)\right\|^2}{\left(\theta_i^{(k)}\right)^2} \tag{45}$$

can be seen as the relative accuracy to which the ith Ritz value approximates an eigenvalue, and the size of the term

$$\frac{\lambda_1}{\theta_i^{(k)} - \lambda_1}, \qquad i \geq 2, \tag{46}$$

depends on the position of $\theta_i^{(k)}$ relatively to the smallest eigenvalue. In particular, one can expect that the term (46) can be of size $\mathcal{O}(1)$ if $\theta_i^{(k)}$ approximates smallest eigenvalues, and it is small if $\theta_i^{(k)}$ approximates largest eigenvalues.

Using the previous simplifications and assuming phase 2, the basic Gauss-Radau upper bound (21) and the rightmost upper bound in (18) are close to each other if and only if

$$\frac{\eta_{1,k}^{(\mu)}}{\mu} + \sum_{i=2}^{k} \left(\frac{\beta_k s_{k,i}^{(k)}}{\theta_i^{(k)}}\right)^2 \frac{\lambda_1}{\theta_i^{(k)} - \lambda_1} \ll \phi_k^{-1}. \tag{47}$$

From Section 5.2 we know that $\eta_{1,k}^{(\mu)}$ goes to zero in phase 2. Hence, if

$$\eta_{1,k}^{(\mu)} < \mu, \tag{48}$$

which will happen for k sufficiently large, then the first term in (47) is smaller than the term on the right-hand side.

As already mentioned, the sum of positive terms in (47) depends only on approximation properties of the underlying Lanczos process, that are not easy to predict in general. Inspired by our model problem described in Section 4, we can just give an intuitive explanation why the sum could be small in phase 2.

Phase 2 occurs in later CG iterations and it is related to the convergence of the smallest Ritz value to the smallest eigenvalue. If the smallest eigenvalue is well approximated by the smallest Ritz value (to a high relative accuracy), then one can expect that many eigenvalues of A are relatively well approximated by Ritz values. If the eigenvalue λ_j of A is well separated from the other eigenvalues and if it is well approximated by a Ritz value, then the corresponding term (45) measuring the relative accuracy to which λ_j is approximated, is going to be small.

In particular, in our model problem, the smallest eigenvalues are well separated from each other, and in phase 2 they are well approximated by Ritz values. Therefore, the corresponding terms (45) are small. Hence, the Ritz values that did not converge yet in phase 2, are going to approximate eigenvalues in clusters which do not correspond to smallest eigenvalues, i.e., for which the terms (46) are small; see also Figs. 3 and 2. In our model problem, the sum of positive terms in (47) is small in phase 2 because either (45) or (46) are small. Therefore, one can expect that the validity of (47) will mainly depend on the size of the first term in (47); see Fig. 7.

The size of the sum of positive terms in (47) obviously depends on the clustering and the distribution of the eigenvalues, and we cannot guarantee in general that it will be small in phase 2. For example, it need not be small if the smallest eigenvalues of A are clustered.
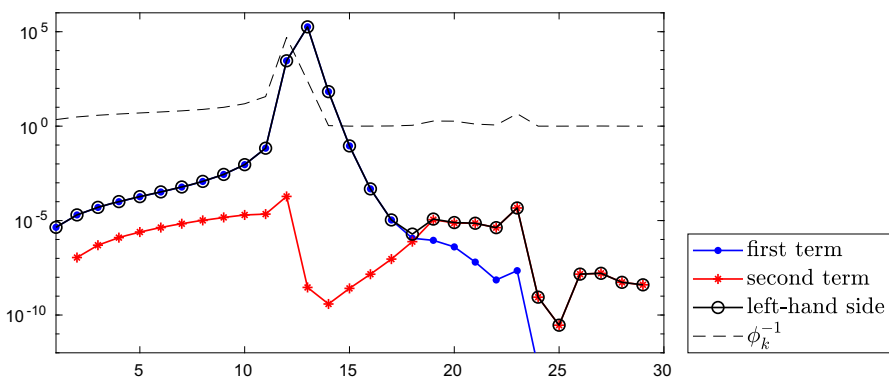

Fig. 7 The first and second term in (44), left-hand side of (43), and ϕ_k^{-1}

7 Detection of phase 2

For our model problem it is not hard to detect phase 2 from the coefficients that are available during the computations. We first observe, see Fig. 7, that the coefficients

$$\gamma_k^{(\mu)} \quad \text{and} \quad \frac{\phi_k}{\mu}, \tag{49}$$

and the corresponding bounds (21) and (18) visually coincide from the beginning up to some iteration ℓ_1. From iteration $\ell_1 + 1$, the Gauss-Radau upper bound (21) starts to be a much better approximation to the squared A-norm of the error than the simple upper bound (18). When phase 2 occurs, the Gauss-Radau upper bound (21) loses its accuracy and, starting from iteration ℓ_2 (approximately when (48) holds), it will again visually coincide with the simple upper bound (18). We observe that phase 2 occurs at some iteration k where the two coefficients (49) significantly differ, i.e., for $\ell_1 < k < \ell_2$. To measure the agreement between the coefficients (49), we can use the easily computable relative distance

$$\frac{\frac{\phi_k}{\mu} - \gamma_k^{(\mu)}}{\gamma_k^{(\mu)}} = \phi_k \left[\left(\frac{\mu}{\theta_1^{(k)}} \right)^2 \frac{\eta_{1,k}^{(\mu)}}{\mu} + \sum_{i=2}^{k} \left(\frac{\beta_k s_{k,i}^{(k)}}{\theta_i^{(k)}} \right)^2 \frac{\mu}{\theta_i^{(k)} - \mu} \right]. \tag{50}$$

We will consider this relative distance to be small, if it is smaller than 0.5.

The behavior of the term in (50) for various values of μ is shown in Fig. 8. The index $\ell_1 = 12$ is the same for all considered values of μ. For μ_3 we get $\ell_2 = 15$ (red circle), for μ_8 we get $\ell_2 = 18$ (magenta circle), for μ_{16} $\ell_2 = 25$ (blue circle), and finally, for μ_{50} there is no index ℓ_2.

As explained in the previous section, in more complicated cases we cannot guarantee in general a similar behavior of the relative distance (50) as in our model problem. For example, in many practical problems we sometimes observe a staircase behavior of the A-norm of the error, when few iterations of stagnation are followed by few iterations of rapid convergence. In such cases, the quantity (50) can oscillate several
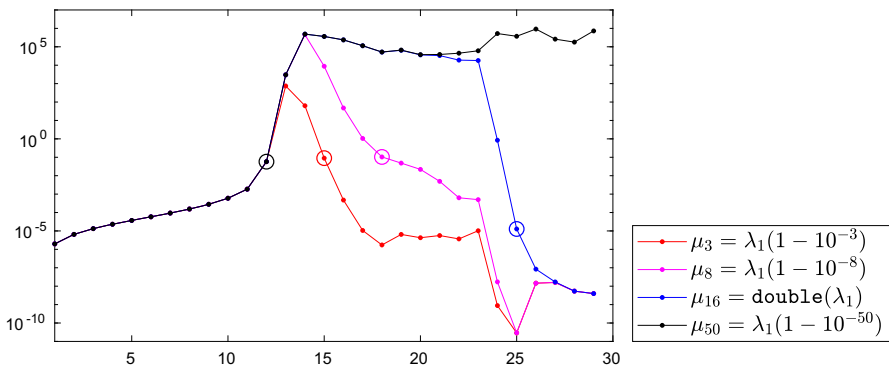

Fig. 8 The behavior of the relative distance in (50) for various values of μ

times and it can be impossible to use it for detecting phase 2. Therefore, in general, we are not able to detect the beginning of phase 2 using (50) reliably. Nevertheless, in particular cases, the formulas (41) and (50) can be helpful.

8 Upper bounds with a guaranteed accuracy

In some applications it might be of interest to obtain upper bounds on the A-norm of the error that are sufficiently accurate. From the previous sections we know that the basic Gauss-Radau upper bound at iteration k can be delayed, and, therefore, it can overestimate the quantity of interest significantly. Nevertheless, going back in the convergence history, we can easily find an iteration index $\ell \leq k$ such that for all $0 \leq i \leq \ell$, the sufficiently accurate upper bound can be found. To find such ℓ, we will use the ideas described in [12] and [19].

For integers $k \geq j \geq \ell \geq 0$, let us denote

$$\Delta_j = \gamma_j \left\| r_j \right\|^2, \quad \Delta_{\ell:k} = \sum_{j=\ell}^{k} \Delta_j, \quad \text{and} \quad \Delta_{j:j-1} = 0.$$

Denoting $\varepsilon_j \equiv \|x - x_j\|_A^2$, the relation (20) takes the form

$$\varepsilon_\ell = \Delta_{\ell:k-1} + \varepsilon_k, \tag{51}$$

A more accurate bound at iteration ℓ is obtained such that the last term in (51) is replaced by the basic lower or upper bounds on ε_k. In particular, the improved Gauss-Radau upper bound at iteration ℓ can be defined as

$$\Omega_{\ell:k}^{(\mu)} = \Delta_{\ell:k-1} + \gamma_k^{(\mu)} \left\| r_k \right\|^2, \tag{52}$$

and the improved Gauss lower bound is given by $\Delta_{\ell:k}$.

To guarantee the relative accuracy of the improved Gauss-Radau upper bound, we would like to find the largest iteration index $\ell \leq k$ in the convergence history such that

$$\frac{\Omega_{\ell:k}^{(\mu)} - \varepsilon_\ell}{\varepsilon_\ell} \leq \tau \tag{53}$$

where τ is a prescribed tolerance, say, $\tau = 0.25$. Since

$$\frac{\Omega_{\ell:k}^{(\mu)} - \varepsilon_\ell}{\varepsilon_\ell} < \frac{\Omega_{\ell:k}^{(\mu)} - \Delta_{\ell:k}}{\Delta_{\ell:k}} = \frac{\|r_k\|^2 \left(\gamma_k^{(\mu)} - \gamma_k \right)}{\Delta_{\ell:k}},$$

we can require $\ell \leq k$ to be the largest integer such that

$$\frac{\|r_k\|^2 \left(\gamma_k^{(\mu)} - \gamma_k \right)}{\Delta_{\ell:k}} \leq \tau. \tag{54}$$

If (54) holds, then also (53) holds. The just described adaptive strategy for obtaining ℓ giving a sufficiently accurate upper bound is summarized in Algorithm 3.

Algorithm 3 CG with the improved Gauss-Radau upper bound.

---

input A, b, x_0, μ, τ
$r_0 = b - Ax_0$, $p_0 = r_0$
$\ell = 0, \gamma_0^{(\mu)} = \frac{1}{\mu}$
for $k = 0, \dots,$ **do**
   cgiter(k)
   while $k \geq \ell$ and (54) **do**
     accept $\Omega_{\ell:k}^{(\mu)}$
     $\ell = \ell + 1$
   end while
   $\gamma_{k+1}^{(\mu)} = \frac{\gamma_k^{(\mu)} - \gamma_k}{\mu(\gamma_k^{(\mu)} - \gamma_k) + \delta_{k+1}}$
end for

Note that

$$\frac{\Omega_{\ell:k}^{(\mu)} - \varepsilon_\ell}{\varepsilon_\ell} + \frac{\varepsilon_\ell - \Delta_{\ell:k}}{\varepsilon_\ell} < \frac{\Omega_{\ell:k}^{(\mu)} - \Delta_{\ell:k}}{\Delta_{\ell:k}},$$

i.e., if (54) holds, then τ represents also an upper bound on the sum of relative errors of the improved lower and upper bounds. In other words, if ℓ is such that (54) is satisfied, then both the improved Gauss-Radau upper bound as well as the improved Gauss lower bound are sufficiently accurate. For a heuristic strategy focused on improving the accuracy of the Gauss lower bound, see [19].

In the previous sections we have seen that the basic Gauss-Radau upper bound is delayed, in particular in phase 2. The delay of the basic Gauss-Radau upper bound can be defined as the smallest nonnegative integer j such that

$$\gamma_{\ell+j+1}^{(\mu)} \left\| r_{\ell+j+1} \right\|^2 < \varepsilon_\ell. \tag{55}$$

Having sufficiently accurate lower and upper bounds (e.g., if (54) is satisfied), we can approximately determine the delay of the basic Gauss-Radau upper bound as the smallest j satisfying (55) where ε_ℓ in (55) is replaced by its tight lower bound $\Delta_{\ell:k}$.

9 Conclusions

In this paper we discussed and analyzed the behavior of the Gauss-Radau upper bound on the A-norm of the error in CG. In particular, we concentrated on the phenomenon observed during computations showing that, in later CG iterations, the upper bound loses its accuracy, it is almost independent of μ, and visually coincides with the simple upper bound. We explained that this phenomenon is closely related to the convergence of the smallest Ritz value to the smallest eigenvalue of A. It occurs when the smallest Ritz value is a better approximation to the smallest eigenvalue

than the prescribed underestimate μ. We developed formulas that can be helpful in understanding this behavior. Note that the loss of accuracy of the Gauss-Radau upper bound is not directly linked to rounding errors in computations of the bound, but it is related to the finite precision behavior of the underlying Lanczos process. In more detail, the phenomenon can occur when solving linear systems with clustered eigenvalues. However, the results of finite precision CG computations can be seen (up to some small inaccuracies) as the results of the exact CG algorithm applied to a larger system with the system matrix having clustered eigenvalues. Therefore, one can expect that the discussed phenomenon can occur in practical computations not only when A has clustered eigenvalues, but also whenever orthogonality is lost in the CG algorithm.

Data Availability All data generated or analyzed during this study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate The authors have read the Springer journal policies on authors responsibilities (including Ethical responsibilities of authors) and submit this manuscript in accordance with those policies.

Consent for publication We have read and understood the publishing policy, and submit this manuscript in accordance with this policy.

Competing interests The authors declare no competing interests.

References

1. Meurant, G., Tichý, P.: Approximating the extreme Ritz values and upper bounds for the A-norm of the error in CG. Numer. Algorithms **82**(3), 937–968 (2019)
2. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Research Nat. Bur. Standards **49**, 409–436 (1952)
3. Dahlquist, G., Eisenstat, S.C., Golub, G.H.: Bounds for the error of linear systems of equations using the theory of moments. J. Math. Anal. Appl. **37**, 151–166 (1972)

4. Dahlquist, G., Golub, G.H., Nash, S.G.: Bounds for the error in linear systems. In: Semi-infinite Programming (Proc. Workshop, Bad Honnef, 1978). Lecture Notes in Control and Information Sci., vol. 15, pp. 154–172. Springer, Berlin (1979)
5. Golub, G.H., Meurant, G.: Matrices, moments and quadrature. In: Numerical Analysis 1993 (Dundee, 1993). Pitman Res. Notes Math. Ser., vol. 303, pp. 105–156. Longman Sci. Tech., Harlow (1994)
6. Golub, G.H., Meurant, G.: Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. BIT **37**(3), 687–705 (1997)
7. Golub, G.H., Strakoš, Z.: Estimates in quadratic formulas. Numer. Algorithms **8**(2–4), 241–268 (1994)
8. Meurant, G.: The computation of bounds for the norm of the error in the conjugate gradient algorithm. Numer. Algo. **16**(1), 77–87 (1998)
9. Meurant, G.: Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. Numer. Algo. **22**(3–4), 353–365 (1999)
10. Meurant, G., Tichý, P.: On computing quadrature-based bounds for the A-norm of the error in conjugate gradients. Numer. Algo. **62**(2), 163–191 (2013)
11. Meurant, G., Tichý, P.: Erratum to: On computing quadrature-based bounds for the A-norm of the error in conjugate gradients [mr3011386]. Numer. Algorithms **66**(3), 679–680 (2014)
12. Strakoš, Z., Tichý, P.: On error estimation in the conjugate gradient method and why it works in finite precision computations. Electron. Trans. Numer. Anal. **13**, 56–80 (2002)
13. Strakoš, Z., Tichý, P.: Error estimation in preconditioned conjugate gradients. BIT **45**(4), 789–817 (2005)
14. Parlett, B.N.: The Symmetric Eigenvalue Problem. Classics in Applied Mathematics, vol. 20, p. 398. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1998). Corrected reprint of the 1980 original
15. Demmel, J.W.: Applied Numerical Linear Algebra, p. 419. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1997)
16. Meurant, G.: The Lanczos and Conjugate Gradient Algorithms. Software, Environments, and Tools, vol. 19, p. 365. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2006)
17. Golub, G.H.: Some modified matrix eigenvalue problems. SIAM Rev. **15**, 318–334 (1973)
18. Golub, G.H., Meurant, G.: Matrices, Moments and Quadrature with Applications, p. 698. Princeton University Press, USA (2010)
19. Meurant, G., Papež, J., Tichý, P.: Accurate error estimation in CG. Numer. Algorithms **88**(3), 1337–1359 (2021)
20. Greenbaum, A.: Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. Linear Algebra Appl. **113**, 7–63 (1989)
21. Greenbaum, A., Strakoš, Z.: Predicting the behavior of finite precision Lanczos and conjugate gradient computations. SIAM J. Matrix Anal. Appl. **13**(1), 121–137 (1992)
22. O'Leary, D.P., Strakoš, Z., Tichý, P.: On sensitivity of Gauss-Christoffel quadrature. Numer. Math. **107**(1), 147–174 (2007)
23. Strakoš, Z.: On the real convergence rate of the conjugate gradient method. Linear Algebra Appl. **154**(156), 535–549 (1991)
24. Paige, C.C.: Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. Linear Algebra Appl. **34**, 235–258 (1980)
25. Gragg, W.B., Harrod, W.J.: The numerically stable reconstruction of Jacobi matrices from spectral data. Numer. Math. **44**(3), 317–335 (1984)
26. Gautschi, W.: Orthogonal Polynomials: Computation and Approximation. Oxford University Press, UK (2004)
27. Scott, D.S.: How to make the Lanczos algorithm converge slowly. Math. Comp. **33**(145), 239–247 (1979)
28. Meurant, G.: On prescribing the convergence behavior of the conjugate gradient algorithm. Numer. Algorithms **84**(4), 1353–1380 (2020)