Numerical Algorithms manuscript No.
(will be inserted by the editor)

Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations

Tomáš Gergelits · Zdeněk Strakoš

the date of receipt and acceptance should be inserted later

Abstract The conjugate gradient method (CG) for solving linear systems of algebraic equations represents a *highly nonlinear finite process*. Since the original paper of Hestenes and Stiefel published in 1952, it has been linked with the Gauss-Christoffel quadrature approximation of Riemann-Stieltjes distribution functions determined by the data, i.e., with a simplified form of the *Stieltjes moment problem*. This link, developed further by Vorobyev, Brezinski, Golub, Meurant and others, indicates that a general description of the CG rate of convergence using an asymptotic convergence factor has principal limitations. Moreover, CG is computationally based on *short recurrences*. In finite precision arithmetic its behaviour is therefore affected by a possible loss of orthogonality among the computed direction vectors. Consequently, *any* consideration concerning the CG rate of convergence relevant to practical computations must include analysis of effects of rounding errors.

Through the example of composite convergence bounds based on Chebyshev polynomials, this paper argues that the facts mentioned above should become a part of common considerations on the CG rate of convergence. It also explains that the spectrum composed of small number of well separated tight clusters of eigenvalues does not necessarily imply a fast convergence of CG or other Krylov subspace methods.

Mathematics Subject Classification (2000) 65F10, 65B99, 65G50, 65N15

The work was supported by the ERC-CZ project LL1202, by the GACR grant 201/09/0917 and by the GAUK grant 695612.

Tomáš Gergelits · Zdeněk Strakoš
Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic
E-mail: gergelits.tomas@seznam.cz · strakos@karlin.mff.cuni.cz

Keywords Conjugate gradient method · Stieltjes moment problem · Chebyshev semi-iterative method · Composite polynomial convergence bounds · Finite precision computations · Clusters of eigenvalues

1 Introduction

In this paper we consider the method of conjugate gradients (CG) [33] for solving linear algebraic systems $Ax = b$, where $A \in \mathbb{C}^{N \times N}$ is Hermitian and positive definite (HPD) matrix which is typically large and sparse. Given an initial guess x_0 and $r_0 = b - Ax_0$, the CG approximations x_k are uniquely determined by the relations

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots,$$

where $r_k = b - Ax_k$ is the k-th residual and

$$\mathcal{K}_k(A, r_0) \equiv \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$$

is the k-th Krylov subspace associated with the matrix A and the vector r_0.

Apart from simple examples, CG can not be applied without *preconditioning*. Throughout this paper we assume that $Ax = b$ represents the preconditioned system. CG can be introduced in more general infinite dimensional Hilbert space settings; see, e.g. [61, Chapter III, Sections 2 and 4], [19,62], and also the recent descriptions using the Riesz map in, e.g., [38,30]. Throughout this paper, the finite dimensional linear algebraic formulation will be sufficient. If A and b results from preconditioning of discretized operator equation (as in numerical solution of partial differential equations), then the preconditioning is often motivated by the operator context; see, e.g. [59,22,5,34,7,52,38]. In practical computations, preconditioning is incorporated into the algorithm and the preconditioned system $Ax = b$ is not formed. For an analytic investigation of the rate of convergence assuming exact arithmetic this difference is not important. In finite precision arithmetic, convergence is delayed due to the loss of orthogonality among the computed direction (residual) vectors. This can be conveniently demonstrated using the preconditioned system $Ax = b$ without going into further details on the particular preconditioning technique. An example of a detailed rounding error analysis can be found, e.g., in [56].

1.1 CG, Gauss-Christoffel quadrature and the Stieltjes moment problem

Throughout the paper we assume that $A \in \mathbb{C}^{N \times N}$ is HPD with the spectral decomposition

$$A = U \,\text{diag}(\lambda_1, \ldots, \lambda_N)\, U^*, \;\; U^*U = UU^* = I \tag{1}$$

where for simplicity of notation $0 < \lambda_1 < \ldots < \lambda_N$ and $U = [u_1, \ldots, u_N]$. Using this spectral decomposition, $v_1 \equiv r_0/\|r_0\|$ and $\omega_j \equiv |(v_1, u_j)|^2, \;\; j =$

$1, \ldots, N$, the moments of the distribution function $\omega(\lambda)$ determined by the nodes $\lambda_1, \ldots, \lambda_N$ and the weights $\omega_1, \ldots, \omega_N$ are given by

$$\sum_{j=1}^{N} \omega_j \lambda_j^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \ldots . \tag{2}$$

The n-node Gauss-Christoffel quadrature of the monomials then determines the n nodes $\theta_l^{(n)}$ and weights $\omega_l^{(n)}$, $l = 1, \ldots, n$, of the distribution function $\omega^{(n)}(\lambda)$ such that the first $2n$ moments of the distribution function $\omega(\lambda)$ are matched, i.e.,

$$\sum_{l=1}^{n} \omega_l^{(n)} \{\theta_j^{(n)}\}^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \ldots, 2n. \tag{3}$$

Here the sums on the left hand sides of (2) and (3) can be expressed via the Riemann-Stieltjes integrals for the monomials with respect to the distribution functions $\omega(\lambda)$ and $\omega^{(n)}(\lambda)$ respectively.

As explained in [37, Section 3.5] with references to many earlier publications, CG applied to $Ax = b$ with the initial residual r_0 can be understood as a process generating the sequence of the distribution functions $\omega^{(n)}(\lambda),\ n = 1, \ldots, N$ approximating the original distribution function $\omega(\lambda)$ in the sense of the Gauss-Christoffel quadrature. Equivalently, CG (implicitly) solves the (simplified) Stieltjes moment problem (2)–(3). The energy norm of the CG error is then given by

$$\begin{aligned} \|x - x_n\|_A^2 =& \|r_0\|^2 \left(\sum_{j=1}^{N} \omega_j \lambda_j^{-1} - \sum_{l=1}^{n} \omega_l^{(n)} \{\theta_l^{(n)}\}^{-1} \right) & (4) \\ =& \|r_0\|^2 \sum_{j=1}^{N} \prod_{l=1}^{n} \left(\frac{1}{\lambda_j^{1/2}} - \frac{\lambda_j^{1/2}}{\theta_l^{(n)}} \right)^2 \omega_j \, ; & (5) \end{aligned}$$

see [37, Section 5.6.1, Corollary 5.6.2 and Theorem 5.6.3]. The nodes $\theta_l^{(n)}$ and the weights $\omega_l^{(n)}$ are the eigenvalues and the squared first components of the associated normalized eigenvectors of the Jacobi matrix T_n generated in the first n steps of the Lanczos process applied to the matrix A with the initial vector v_1. The matrix T_n represents the *operator* $A : \mathbb{C}^N \to \mathbb{C}^N$ restricted and orthogonally projected onto the n-th Krylov subspace $\mathcal{K}_n(A, r_0)$, which reveals the degree of nonlinearity with respect to A; see, e.g., [61,12,37][1].

Recalling the previous facts prior to starting a discussion of a-priori bounds or estimates for the CG rate of convergence (based on some simplified information extracted from A and b) makes a good sense for the following reason. *Any* such bound or estimate has to deal with the tremendous *nonlinear complexity* of the expressions (4) and (5). Further details can be found, e.g., in [37,24,41].

[1] The nonlinearity with respect to b has recently been studied in [26].

1.2 Comments on the a-priori analysis of the CG rate of convergence

A-priori analysis of the rate of convergence of CG (as well as of other Krylov subspace methods) focuses on certain relatively simple characteristics of the problem which can conveniently be linked (if applicable) with the underlying system of infinite dimensional operator equations, its preconditioning and discretisation. A condition number of the preconditioned discretized operator in combination with some information on large or small eigenvalues may serve as the most typical example of such characteristics. Following the functional analysis-based investigation in [45] as well as experimental observations, it is *assumed* that the rate of convergence follows the following three consecutive phases (see [45, Section 1.3]):

> "*in the early sweeps the convergence is very rapid but then slows down, this is the sublinear behavior. The convergence then settles down to a roughly constant linear rate. ... Towards the end new speed may be picked up again, corresponding to the superlinear behavior.*"

Heuristic arguments on CG based on the spectrum of A are used to support this assumption (see also [6, Section 1]). It should be taken into account, however, that this assumption and the supporting heuristics are based on experience with *some* spectral distributions. It can not be generalized to all practical problems. This is made clear in [45] by the sentence almost immediately following the quoted one given above:

> "*In practice all phases need not be identifiable, nor they appear only once and in this order.*"

The sublinear, linear and superlinear phases are analysed in literature using various tools; see, e.g., [45,62,6] or the survey in [7, Sections 2–4]. Section 3.2 of [6] gives a nice example on how the reasoning about an initial sublinear phase can be applied in practice; see also [4].

Applications of the results associated with particular phases to practical computations or to analysis of a particular problem requires verification whether the assumptions used in derivations are met in the given problems. Here the asymptotic reasoning requires a special attention. As stated in [22, p. 113]:

> "*Methods with similar asymptotic work estimates may behave quite differently in practice*".

Krylov subspace methods are mathematically *finite*. Therefore, strictly speaking, in Krylov subspace methods there is no asymptotic present at all.

In relation to the last point it is sometimes argued in literature that due to rounding errors Krylov subspace methods do not terminate in a finite number of steps and *therefore* they are considered iterative methods which also justifies use of asymptotic bounds. In our opinion this point is not valid. First, effects of rounding errors depend on whether methods are implemented via short or long recurrences; see [37, Sections 5.9 and 5.10]. Second, the standard

CG implementation is based (for a good reason; see, e.g., the surveys in [41] and [31]) on coupled two-term recurrences. In finite precision arithmetic the orthogonality of the computed residuals (or direction vectors) can not be, in general, preserved, which results in a *delay of convergence*. The mechanism of this delay is well understood, and its consequences should not be interpreted as making the iteration process infinite.

This is immediately clear from the other effect of rounding errors, called *maximal attainable accuracy*. The accuracy of the computed approximate solution can not be improved below some level of the error determined by the implementation, computer arithmetic and the input data; see, e.g., [28, Section 7.3], [41, Section 5.4], [37, Section 5.9.3] and the references given there. CG as well as other Krylov subspace methods are considered iterative because the iteration can be stopped whenever the user-specified accuracy is reached; see, e.g. [32, Section 2.4.2] and [3, p. 450]. The stopping criteria must be based on *a-posteriori* error analysis; see, e.g. [1, in particular Section 4.1] for a recent survey of the context in adaptive numerical solution of elliptic partial differential equations, as well as [20] and [3, Appendix A] for some early examples.

Throughout this paper we *assume* that the iteration is stopped before the maximal attainable accuracy is reached. Such assumption can not be taken in practical computations for granted. It must be justified by a proper numerical stability analysis (a simple *a-posteriori* check can be based on comparison of the iteratively and directly computed residuals). A detailed exposition of the related issues is out of the scope of this paper and we refer the interested reader to the literature given above.

In summary, *a-priori* analysis of the CG rate of convergence must take into account a possible delay of convergence due to rounding errors. Since in CG computations keeping short recurrences is essential, which inevitably results in a loss of orthogonality, *developing bounds or estimates which are to be applied to practical computations can not assume exact arithmetic.*

1.3 Analysis based on Chebyshev polynomials

In this paper we focus on the most common *a-priori* analysis of the CG convergence rate based on Chebyshev polynomials. The rate of convergence of CG is associated with linear convergence bounds derived using scaled and shifted Chebyshev polynomials in hundreds of papers and essentially in every textbook covering the CG method. As argued in Section 1.1 above, the CG method and therefore also its convergence rate are, however, nonlinear and its convergence often tends to accelerate, with more or less pronounced variations, during the iteration process. Axelsson [2] and Jennings [35] suggested in this context composite polynomial bounds based on explicit annihilation of the outlying eigenvalues. Such bounds seemed to offer an illustrative explanation especially in case when large outlying eigenvalues were present in the spectrum.[2] These composite polynomial bounds assumed exact arithmetic. As rounding errors may substantially delay convergence of the CG method, it is not clear whether the composite polynomial bounds and the conclusions based on them apply to finite precision CG computations. A motivating example is presented in Figure 1. It indeed shows that a composite polynomial bound can fail to describe CG convergence quantitatively and even *qualitatively.* The difficulty has been to some extent noticed already by Jennings in the paper [35], and also by van der Sluis and van der Vorst [58] who therefore restrict themselves to the case of small outlying eigenvalues, where the difficulty caused by finite precision arithmetic is not strongly pronounced. In the rest of the paper we deal with the composite polynomial bounds with large outlying eigenvalues. They are used for quantitative evaluation of CG convergence and conclusions based on them are published in recent literature.

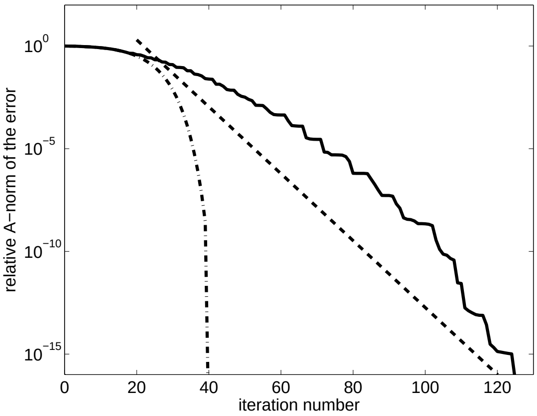

Fig. 1 Rounding errors can cause a substantial delay of convergence in finite precision CG computations (solid line) in comparison to their exact precision counterpart (dash-dotted line). A composite polynomial bound (dashed line) fails to describe the finite precision CG behaviour *quantitatively* (the slope given by the bound is not descriptive) and *qualitatively* (the staircase-like shape of the convergence curve).

The paper is organized as follows. In Section 2 we briefly clarify the relationship between the CG method, the CSI method and the well known linear convergence bound derived using Chebyshev polynomials. Section 3 describes the construction of the composite polynomial bounds and comments on their properties. In Section 4 we use results of the backward-like analysis by Greenbaum and compare *exact* CG computations where matrices have well separated individual eigenvalues, with exact CG computations where matrices have corresponding well separated *clusters* of eigenvalues. We conclude that a "bird's eye view" of the spectrum can be misleading in Krylov subspace methods.

[2] It should be understood, however, that the spectral upper bounds do not necessarily describe the actual CG convergence behaviour for particular right hand sides (initial residuals); see, e.g., [37, Sections 5.6.1–5.6.3] and [9–11, 43, 44].

Based on that we examine validity of the composite polynomial bounds for finite precision CG computations. We conclude and numerically demonstrate that in the presence of large outlying eigenvalues such bounds have, apart from simple exceptions, little in common with the finite precision behaviour of the CG method. Section 5 presents numerical experiments which illustrate in detail shortcomings of the composite polynomial bounds. Concluding remarks summarize the presented clarifications and formulate recommendations for evaluation of the CG rate of convergence.

Writing this paper is motivated by persisting misunderstandings reappearing in literature. This is not meant as a criticism or a negative statement. Our point is that the whole matter is very complex and this should be taken into account whenever any simplification is made. The presented formulas are not new, but, except for the Chapter 5 of the monograph [37], they have not been, to our knowledge, presented in a comprehensive way in a single publication. Most of the points are presented in [37], but their placement is subordinate to the organization of the whole monograph, which addresses many related as well as many distant topics. Therefore we consider useful to publish this focused presentation, which in some parts (in particular Section 4 and Section 5) complements the presentation in [37] by some new observations. In comparison to a monograph covering much larger area, presentation in the paper allows to focus on interpretation of the formulas. We believe that here the interpretation is more important than the formulas themselves. A need for the correct interpretation can be underlined by the following quote presented (in a somewhat related context) in the instructive paper by Faber, Manteuffel and Parter [22, p. 113]:

> “There is no flaw in the analysis, only a flaw in the conclusions drawn from the analysis.”

2 Chebyshev semi-iterative method, conjugate gradient method and their convergence bounds

The idea of the Chebyshev semi-iterative (CSI) method can be linked, with the works of Flanders and Shortley [23], Lanczos [36] and Young [63]. The CSI method requires a knowledge or estimation of the extreme eigenvalues $\lambda_1 < \lambda_N$ of A and it can be implemented using the three-term recurrence relation for the Chebyshev polynomials; see, e.g., [60, Chapter 5].

The CSI method can be viewed as a polynomial acceleration of the stationary Richardson iterations [50] where the k-th error can be written as

$$x - x_k = \phi_k^R(A)(x - x_0), \tag{6}$$

and the iteration polynomial

$$\phi_k^R(\lambda) = \left(1 - \frac{2\lambda}{\lambda_1 + \lambda_N}\right)^k$$

belongs to the set of polynomials of degree k with the constant term equal to one (i.e. having the value one at zero). As has been already observed by Richardson in [50], replacing the k-multiple root of the iteration polynomial $\phi_k^R(\lambda)$ by k distinct roots may lead to faster convergence. The CSI method is motivated by the following reasoning. Let

$$x - x_k = \phi_k(A)(x - x_0),$$

where $\phi_k(\lambda),\ \phi_k(0) = 1$, represents the polynomial of degree at most k. Then the A-norm of the error

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

is given by

$$\|x - x_k\|_A = \|\phi_k(A)(x - x_0)\|_A \tag{7}$$

and using the spectral decomposition (1) of A the relative A-norm of the error satisfies

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le \|\phi_k(A)\| = \max_{j=1,\ldots,N} |\phi_k(\lambda_j)|\,. \tag{8}$$

The right hand side in (8) is independent of the right hand side b and thus it represents the worst case upper bound. Maximizing over the whole interval $[\lambda_1, \lambda_N]$ instead of the discrete set of eigenvalues $\lambda_1, \ldots, \lambda_N$ gives the bound

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le \max_{\lambda \in [\lambda_1, \lambda_N]} |\phi_k(\lambda)|\,. \tag{9}$$

Setting the roots of the iteration polynomial $\phi_k(\lambda)$ as the roots of the shifted Chebyshev polynomial

$$\chi_k(\lambda) = \begin{cases} \cos\left(k \arccos\left(\dfrac{2\lambda - \lambda_N - \lambda_1}{\lambda_N - \lambda_1}\right)\right) & \text{for } \lambda \in [\lambda_1, \lambda_N], \\ \cosh\left(k \operatorname{arccosh}\left(\dfrac{2\lambda - \lambda_N - \lambda_1}{\lambda_N - \lambda_1}\right)\right) & \text{for } \lambda \notin [\lambda_1, \lambda_N], \end{cases} \tag{10}$$

is motivated by the fact that

$$\phi_k(\lambda) \equiv \chi_k(\lambda)/\chi_k(0) \tag{11}$$

represents the unique solution of the minimization problem

$$\min_{\substack{\phi(0)=1 \\ \deg(\phi)\le k}} \max_{\lambda \in [\lambda_1, \lambda_N]} |\phi(\lambda)| \tag{12}$$

originally solved by Markov [39]. In words, the k-th shifted and scaled Chebyshev polynomial has the minimal maximum norm on the interval $[\lambda_1, \lambda_N]$ among the set of all polynomials of degree at most k having the value one at zero.

Substituting (11) into (9) and using $|\chi_k(\lambda)| \le 1$ for $\lambda \in [\lambda_1, \lambda_N]$ results in the bound for the relative A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le |\chi_k(0)|^{-1}, \quad k = 0, 1, 2, \dots; \tag{13}$$

see [63, Section 2]. The alternative definition of the Chebyshev polynomials

$$\chi_k(\gamma) = \frac{1}{2}\left(\left(\gamma + (\gamma^2 - 1)^{\frac{1}{2}}\right)^k + \left(\gamma + (\gamma^2 - 1)^{\frac{1}{2}}\right)^{-k}\right) \tag{14}$$

(see, e.g., [51, Section 1.1]) gives with the shift $\gamma = (2\lambda - \lambda_N - \lambda_1)/(\lambda_N - \lambda_1)$ used in (10) after a simple manipulation

$$|\chi_k(0)| = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1}\right)^k + \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k\right] \ge \frac{1}{2}\left(\frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1}\right)^k \tag{15}$$

where $\kappa(A) = \lambda_N/\lambda_1$ is the condition number of A. This gives the convergence bound *for the CSI method*, which was published in this form by Rutishauser [21, II.23] in 1959,

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k, \quad k = 0, 1, 2, \dots. \tag{16}$$

The CG approximations x_k minimize the A-norm of the error over the manifolds $x_0 + \mathcal{K}_k(A, r_0)$; cf. [33, Theorem 4.1]. Equivalently,

$$\begin{aligned} \|x - x_k\|_A &= \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \|\varphi(A)(x - x_0)\|_A & (17)\\ &= \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \left\{\sum_{j=1}^{N} |\xi_j|^2 \lambda_j \varphi^2(\lambda_j)\right\}^{1/2}, & (18) \end{aligned}$$

where $|\xi_j|$ represents the size of the component of the initial error $x - x_0$ in the direction of the eigenvector u_j corresponding to λ_j, i.e.,

$$x - x_0 = \sum_{j=1}^{N} \xi_j u_j \tag{19}$$

and, similarly to (18),

$$\|x - x_0\|_A = \left\{\sum_{j=1}^{N} |\xi_j|^2 \lambda_j\right\}^{1/2}. \tag{20}$$

The formula (17) leads, using the spectral decomposition (1) of A, to the bound for the relative A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \, ; \tag{21}$$

cf. (8). This bound is independent on the right-hand side b and thus it represents the worst case upper bound for the CG method. Since

$$\min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \leq |\chi_k(0)|^{-1} \max_{j=1,\ldots,N} |\chi_k(\lambda_j)| \leq |\chi_k(0)|^{-1} \, , \tag{22}$$

we can apply (15) and conclude that the bound (16) must also hold for the CG method.

Now we come to the point which is fundamental but still very rarely mentioned in literature. It should be acknowledged that (16) *represents the bound for the CSI method*; see the very clear description given by Rutishauser in [21]. This bound holds for the CG method because the optimal polynomial giving the minimum in (21) can be bounded using (22). The behaviour of $\|x - x_k\|_A$ for some given initial error (residual) is, however, given by (18), which can be substantially different than suggested by (21) and therefore certainly substantially different than suggested by the CSI error bound (16). The different nature of the CG and CSI methods is clear also from the comparison of the minimization problems (12) and (18). Whereas the CSI norm of the error can be *tightly* bounded by the minimization problem over the whole interval $[\lambda_1, \lambda_N]$, the CG norm of the error is determined by the discrete minimization problem.

We have presented the (known) derivation in detail in order to avoid further misinterpretations of the relationship between the CSI and CG methods and of the relationship of the bound (16) to the CG rate of convergence. In short, as described in Section 1.1, CG solves the simplified Stieltjes moment problem. Therefore the CG iteration polynomials $\varphi_k(\lambda),\ k = 0, 1, \ldots, N$ defined by (17) are orthogonal with respect to the (discrete) inner product determined by the Riemann-Stieltjes integral with the distribution function $\omega(\lambda)$. The Chebyshev polynomials $\chi_k(\lambda),\ k = 0, 1, \ldots$ are orthogonal with respect to the certain continuous and discrete inner products which contain apart from the extremal eigenvalues λ_1 and λ_N no further information about the data A, b and r_0 (or $x - x_0$); see, e.g. [51, Section 1.5] and [16, Theorem 4.5.20]. Polynomials orthogonal with different inner products can indeed be very different. Therefore it is beyond any doubt that, except for very special situations, *the bound* (16) *relevant for the CSI method has a very little in common with the rate of convergence of the CG method.* Further details and extensive historical comments can be found in [37, Section 5.6.2].

The upper bound (16) implies that, in *exact arithmetic*,

$$k_\epsilon = \left\lceil \frac{1}{2} \ln\left(\frac{2}{\epsilon}\right) \sqrt{\kappa(A)} \right\rceil \tag{23}$$

iterations ensure the decrease of the relative energy norm of the CSI (and therefore also CG) error below the given level of accuracy $\epsilon > 0$ (here $\lceil\cdot\rceil$ denotes rounding up to the nearest integer). As justified in [27,29], using results of a thorough analysis, the presented results hold, with a small correction, also for *finite precision arithmetic* CG computations. When $\kappa(A) = \lambda_N/\lambda_1 \approx 1$, the linear system is easily solvable. Using the bound (16) and the iteration count (23) for CG computations then does not cause any harm. But in such cases one should also ask whether the CG method is really needed for solving such problems. Simpler methods might be fast enough. If $\kappa(A) \gg 1$, then, depending on the *distribution of the spectrum inside the interval* $[\lambda_1, \lambda_N]$, the CG method and the CSI method can naturally perform very differently. In such cases an application of the bound (16) to the CG method should always be accompanied with an appropriate justification.

3 Composite polynomial bounds and superlinear convergence assuming exact arithmetic

As mentioned above, the superlinear convergence behaviour of the CG method in *exact arithmetic* was explained by Axelsson [2] and Jennings [35] using composite polynomial bounds. For any given polynomial $q_m(\lambda)$ of degree $m \leq k$ satisfying $q_m(0) = 1$ we obtain

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \leq \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k-m}} \max_{j=1,\ldots,N} |q_m(\lambda_j)\varphi(\lambda_j)|\,, \tag{24}$$

where the minimax problem on the right hand side considers the composite polynomial $q_m(\lambda)\varphi(\lambda)$. In order to describe the superlinear convergence in case of large outlying eigenvalues, Axelsson and Jennings propose in [2,35] the following natural choice

$$q_m(\lambda) = \prod_{j=N-m+1}^{N} \left(1 - \frac{\lambda}{\lambda_j}\right). \tag{25}$$

Since the polynomial $q_m(\lambda)$ given by (25) has by construction its roots at the m largest eigenvalues, the relative A-norm of the error is bounded, using (21) and (24), as

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k-m}} \max_{j=1,\ldots,N} |q_m(\lambda_j)\varphi(\lambda_j)| \tag{26}$$

$$\leq \max_{j=1,\ldots,N-m} |q_m(\lambda_j)| \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k-m}} \max_{j=1,\ldots,N-m} |\varphi(\lambda_j)|\,. \tag{27}$$

The polynomial $\varphi(\lambda)$ is evaluated only at the eigenvalues $\lambda_1, \ldots, \lambda_{N-m}$. Therefore the use of the composite polynomial

$$q_m(\lambda)\chi_{k-m}(\lambda)/\chi_{k-m}(0), \tag{28}$$

where $\chi_{k-m}(\lambda)$ denotes the Chebyshev polynomial of degree $k-m$ shifted to the interval $[\lambda_1, \lambda_{N-m}]$, results using $|q_m(\lambda_j)| \leq 1$ for $j = 1, \ldots, N-m$, analogously to Section 2, in the bound

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq 2\left(\frac{\sqrt{\kappa_m(A)} - 1}{\sqrt{\kappa_m(A)} + 1}\right)^{k-m}, \quad k = m, m+1, \ldots, \tag{29}$$

where $\kappa_m(A) \equiv \lambda_{N-m}/\lambda_1$ is the so-called *effective condition number*. This quantity is typically substantially smaller than the condition number $\kappa(A)$ which indicates possibly faster convergence after m initial iterations. Illustration of the composite polynomial (28) is for $k = 8$, $m = 2$, and the eigenvalues $\lambda_1 = 0.1, \lambda_{N-2} = 6, \lambda_{N-1} = 9$ and $\lambda_N = 15$ given in Figure 2. As we can immediately observe, the composite polynomial has even for small N, k and small $\kappa(A)$ and $\kappa_m(A)$ very large gradients close to the outlying eigenvalues λ_{N-1} and λ_N. This observation will be important below.

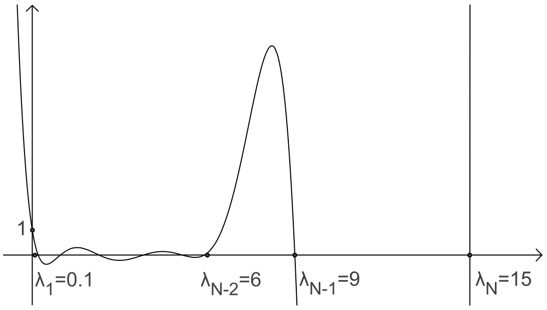

Fig. 2 Illustration of the composite polynomial (28) with $k = 8$ and $m = 2$. The polynomial has roots at two large outlying eigenvalues and on the rest of the spectrum is small due to the minimax property of the Chebyshev polynomials. Here the underlying matrix of dimension N would have two largest eigenvalues $\lambda_N = 15, \lambda_{N-1} = 9$ and the remaining eigenvalues would be arbitrarily distributed in the interval $[0.1, 6]$.

Using an idea analogous to [58], CG computations with the initial error $x - x_0$ are compared in [37, Theorem 5.6.9] to CG computations with the initial error $x - \tilde{x}_0$ obtained from $x - x_0$ by neglecting the components ξ_j in the direction of the m eigenvectors corresponding to the m largest eigenvalues,

$$\|x - \tilde{x}_0\|_A = \left\{\sum_{j=1}^{N-m} |\xi_j|^2 \lambda_j\right\}^{1/2}. \tag{30}$$

This comparison gives the following formula

$$\|x - \tilde{x}_k\|_A \leq \|x - x_k\|_A \leq \|x - \tilde{x}_{k-m}\|_A, \quad k = m, m+1, \ldots \tag{31}$$

The right inequality in (31) shows that CG computation for $Ax = b$ with the initial error $x - x_0$ (the initial residual $r_0 = b - Ax_0$) is from its m-th iteration at least as fast as CG computations for $Ax = b$ with the initial error $x - \tilde{x}_0$ from the start. Dividing this inequality by $\|x - x_0\|_A$ and using $\|x - \tilde{x}_0\|_A \leq \|x - x_0\|_A$ we get the upper bound (29) based on the idea of composite polynomial, indeed

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \frac{\|x - \tilde{x}_{k-m}\|_A}{\|x - \tilde{x}_0\|_A} \leq 2 \left(\frac{\sqrt{\kappa_m(A)} - 1}{\sqrt{\kappa_m(A)} + 1} \right)^{k-m}, \quad k = m, m+1, \ldots. \tag{32}$$

This upper bound can be interpreted as if the first m CG iterations "annihilate" the m large outlying eigenvalues with the subsequent convergence rate bounded linearly by (32). It should be noted, however, that this is nothing but an *interpretation.* CG computations do not work that way; see also [37, Section 5.6.4].

Analogously to (23) in Section 2 we get from the upper bound (32) that after

$$k_\epsilon = m + \left\lceil \frac{1}{2} \ln\left(\frac{2}{\epsilon}\right) \sqrt{\kappa_m(A)} \right\rceil \tag{33}$$

iterations, the relative A-norm of the error drops below the given tolerance ϵ; see [2, p. 132], [35, relation (5.9)] as well as the recent application of this formula in [53, Theorem 2.5].

It should be emphasized, however, that all this is true only in exact arithmetic. The rest of the paper explains that, in general, this approach *must fail in finite precision arithmetic.* The failure of the composite polynomial bounds in finite precision CG computations can be explained by the fact that the closely related Lanczos method computes in finite precision arithmetic repeated approximations of large outlying eigenvalues. This was observed by many authors and it led to results explaining finite precision behaviour of the Lanczos and CG methods; see, in particular, [49,27,40] and the survey [41] referring to extensive further literature. Despite the theoretical and experimental counterarguments, the composite polynomial bounds and the related asymptotic convergence factor ideas with neglecting eigenvalues away from the rest of the spectrum as insignificant are tempting to be used for justification of cost in CG computations; see e.g. [38, Remark 2.1], [57, Section 20.4] and [53, Theorem 2.5]. In the rest of the paper we restrict ourselves to investigation of the bound (32) and the formula (33). Other approaches not based on Chebyshev polynomials should be in the presence of large outlying eigenvalues examined analogously.

4 Analysis of the composite polynomial bounds in finite precision arithmetic

The CG method determines in exact arithmetic an orthogonal basis of the Krylov subspace $\mathcal{K}_k(A, r_0)$ given by the residuals r_j, $j = 0, 1, \ldots, k-1$. However, in finite precision CG computations the orthogonality of the computed residual vectors is (usually quickly) lost and they often become even (numerically) linearly dependent. Consequently, the computed residual vectors may at the step k span a subspace of dimension smaller than k. This *rank-deficiency* of computed Krylov subspace bases thus determines *delay* of convergence of finite precision computations, which can be defined as the difference between the number of iterations required to attain a prescribed accuracy in finite precision computations and the number of iterations required to attain the same accuracy assuming exact arithmetic.

The bound (29) and the number of iterations (33) were derived assuming exact arithmetic and therefore they do not reflect possible delay of convergence. In finite precision CG computations they suffer from a fundamental difficulty outlined in Figure 1 and illustrated in more detail in Figure 3. Here the dashed lines plot the sequence of the composite polynomial bounds (29) with increasing number of the large eigenvalues of A considered as outliers ($m = 0, 3, 6, \ldots$). The bold solid line represents the convergence curve

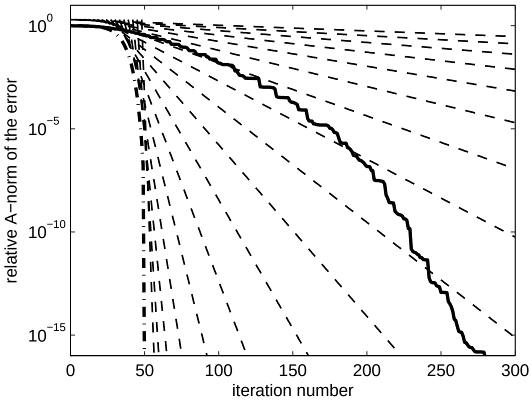

Fig. 3 The sequence of the composite polynomial bounds (29) (dashed lines) for increasing number of deflated large eigenvalues ($m = 0, 3, 6, \ldots$) is compared with the results of finite precision CG computations (bold solid line) and exact CG computations (dash-dotted line).

of the finite precision CG and the dash-dotted line the CG behaviour assuming exact arithmetic[3]. Computations were performed using a symmetric positive definite diagonal matrix A of the size $N = 50$ with the eigenvalues $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_{N-1} < \lambda_N$, where $\lambda_1 = 0.1$, $\lambda_N = 10^4$, the inner eigenvalues were given by the formula

$$\lambda_i = \lambda_1 + \frac{i-1}{N-1}(\lambda_N - \lambda_1)\rho^{N-i} \quad i = 2, \ldots, N-1 \tag{34}$$

and $\rho = 0.8$; see [54,29,40]. The parameter $\rho \in (0, 1]$ determines the non-uniformity of the spectrum. For $\rho \ll 1$ the eigenvalues tend to cumulate near

[3] CG behaviour assuming exact arithmetic is simulated throughout the paper by double reorthogonalization of the residual vectors; see [29,48].

λ_1 and for $\rho = 1$ the spectrum is distributed uniformly. In our experiments we use the vector b of all ones, i.e., $b = [1, \ldots, 1]^T$. We observe that the linear convergence bounds determine (a close) envelope for the exact arithmetic CG convergence curve. This is in correspondence with the intuitive explanation of the superlinear convergence behaviour of CG in exact arithmetic presented in literature. The data in this example do not represent a purely academic case. Spectra with large outlying eigenvalues do appear in practice; see e.g., [8] for an early study on this related to preconditioning techniques.

The point is that *none* of the straight lines describes the finite precision convergence behaviour, as can be seen by comparing the dashed lines with the bold solid line. Evidently, the composite polynomial bounds (29) can not be used, in general, as upper bounds.

The finite precision behaviour of the Lanczos and CG methods was analyzed, in particular, by Paige and Greenbaum; see [27, 48, 49]. Shortly speaking, Greenbaum has proved that

> the finite precision Lanczos computation for a matrix A and a given starting vector v produces in steps 1 through k the same eigenvalue approximations (the same Jacobi matrix T_k) as the exact Lanczos computation for some particular larger matrix $\widehat{A}(k)$ and some particular starting vector $\widehat{v}(k)$ while the eigenvalues of $\widehat{A}(k)$ all lie *within tiny intervals around the eigenvalues of* A. The size as well as (all) individual entries of $\widehat{A}(k)$ and $\widehat{v}(k)$ depend on the rounding errors in the steps 1 through k.

It should be emphasized that $\widehat{A}(k)$ *is not given by a slight perturbation of* A*,* as sometimes stated in literature; $\widehat{A}(k)$ is typically much larger than A. This is illustrated on Figure 4. An analogous statement is valid, with a small inaccuracy specified in [27], also for the behaviour of finite precision CG computations. This explains why (29) and (33) must fail, in general, in finite precision arithmetic, where m CG steps are not enough to annihilate the influence of the m large outlying eigenvalues. One may suggest to resolve the matter by adding several penalty steps which account for the effects of rounding errors. The number of such additional steps, however, depends on current iteration k and it can not be determined a-priori. The difficulty is illustrated in Figure 1 above where the “penalty” is given by the horizontal differences between the dashed line (the bound) and the solid line (computed results).

As stated above, the matrix $\widehat{A}(k)$ and the vector $\widehat{v}(k)$ depend on the iteration step k. The reasoning about the delay in finite precision CG computations suggests (it was experimentally confirmed in [29]) that the particular matrix $\widehat{A}(k)$ constructed for the k steps of the given finite precision CG computation can be replaced (with an acceptable inaccuracy) by a matrix $\widehat{A}$ having sufficiently many eigenvalues in *tight clusters around each eigenvalue of* A; see also the detailed argumentation in [41] and, in particular, in [37, Section 5.9]. The appropriate starting vector associated with $\widehat{A}$ can be constructed from A and b independently of k. As an example, the matrix $\widehat{A}$ used in our experi-

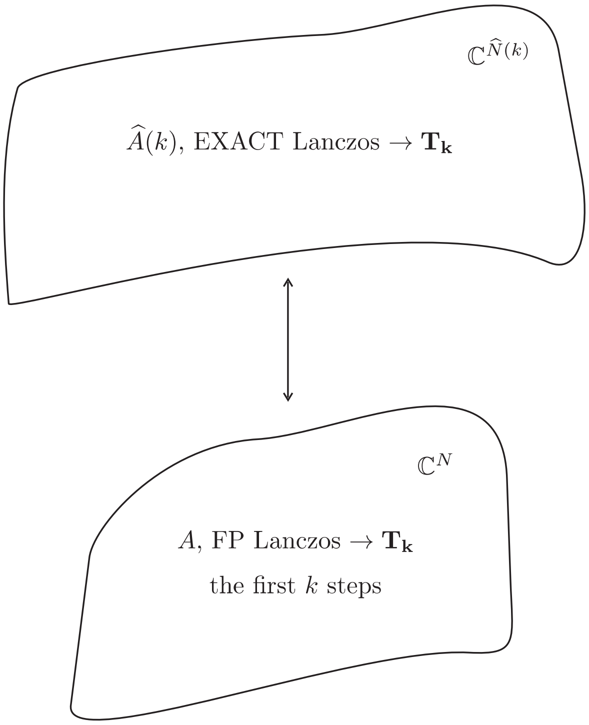
$\mathbb{C}^{\widehat{N}(k)}$
$\widehat{A}(k)$, EXACT Lanczos → $\mathbf{T_k}$
$\mathbb{C}^N$
A, FP Lanczos → $\mathbf{T_k}$
the first k steps

Fig. 4 For any $k = 1, 2, \ldots$ the first k steps of the finite precision Lanczos computation for $A \in \mathbb{C}^{N \times N}$ can be analyzed as the first k steps of the exact Lanczos for the (possibly much larger) matrix $\widehat{A}(k) \in \mathbb{C}^{\widehat{N}(k) \times \widehat{N}(k)}$ depending on k which generates the same $k \times k$ Jacobi matrix T_k.

ments below has l eigenvalues $\widehat{\lambda}_{j,1} < \widehat{\lambda}_{j,2} < \ldots < \widehat{\lambda}_{j,l}$ uniformly distributed in tiny intervals $[\lambda_j - \Delta, \lambda_j + \Delta]$ around each original eigenvalue λ_j of A, $j = 1, 2, \ldots, N$, where l is sufficiently large in correspondence to the maximal number of the performed iterations steps. The associated right hand side $\widehat{b}$ is obtained from b by splitting each individual entry β_j of b into l equal parts $\widehat{\beta}_{j,1}, \ldots, \widehat{\beta}_{j,l}$ such that $\sum_{s=1}^{l} \widehat{\beta}_{j,s}^2 = \beta_j^2$, $j = 1, 2, \ldots, N$; see [29].

As an immediate consequence of the results from [27,29] we get that convergence behaviour of exact CG applied to a matrix with the spectrum having well separated clusters of eigenvalues is both *qualitatively* and *quantitatively* different from the convergence behaviour of *exact* CG applied to a matrix with a spectrum where each cluster is replaced by a single eigenvalue. We can conclude that even for the CG method, the HPD matrix and *assuming exact arithmetic*,

> a spectrum composed of a small number of tight clusters can not be associated, in general, with fast convergence.

Indeed, the associated Stieltjes moment problems from Section 1.1 can be for different distribution of eigenvalues very different. This is true, in particular, when clusters of eigenvalues are replaced by single (representing) eigenvalues of the same weights; see [47]. This point contradicts the common belief which seems widespread.

We will now explain how this fact is reflected in the composite polynomial convergence bounds (29). Using the relationship with the exact CG computations applied to $\widehat{A}$, the corresponding minimization problem which bounds the CG convergence behaviour *in finite precision arithmetic* is not

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \,, \tag{35}$$

where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A; see (21). Instead, one must use

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \max_{\lambda\in\sigma(\widehat{A})} |\varphi(\lambda)| \,, \tag{36}$$

where the spectrum of the matrix $\widehat{A}$ consists of the union of the individual clusters around the original eigenvalues $\lambda_j,\ j = 1, \ldots, N$, i.e., in our case

$$\sigma(\widehat{A}) \equiv \bigcup_{j=1,\ldots,N} \left\{ \widehat{\lambda}_{j,1}, \ldots, \widehat{\lambda}_{j,l} \right\}. \tag{37}$$

Consequently, in order to be valid for finite precision CG computations, the upper bound based on the composite polynomial (28) from Section 3 must use instead of

$$\max_{j=1,\ldots,N} |q_m(\lambda_j)\chi_{k-m}(\lambda_j)| \,/\, |\chi_{k-m}(0)| \,, \tag{38}$$

which considers the values of the composite polynomial at the eigenvalues $\lambda_1, \ldots, \lambda_N$ of A, the modification

$$\max_{\lambda\in\sigma(\widehat{A})} |q_m(\lambda)\chi_{k-m}(\lambda)| \,/\, |\chi_{k-m}(0)| \,, \tag{39}$$

which considers the values of the composite polynomial at the eigenvalues of the matrix $\widehat{A}$. As a consequence of the minimality property of the Chebyshev polynomial $\chi_{k-m}(\lambda)$ over the interval $[\lambda_1, \lambda_{N-m}]$, its values outside this interval become even for small k very large. More specifically, the Chebyshev polynomial is outside the minimality interval the fastest growing polynomial of the given degree; see, e.g., [51, Section 2.7, rel. (2.37)] and [16, Section 3.2.3]. The composite polynomial has, by construction, large values of its gradient at the large outlying eigenvalues of A; see the illustration in Figure 2 above. The values of the composite polynomial at the points located in the tight clusters around such large outlying eigenvalues can therefore be huge even for small k, and the upper bound based on the expression (39) becomes after several iterations in practical computations meaningless; see the illustration in Figure 5. The left part shows finite precision CG convergence behaviour (bold solid line) corresponding to the right hand side b of ones and the matrix A of dimension

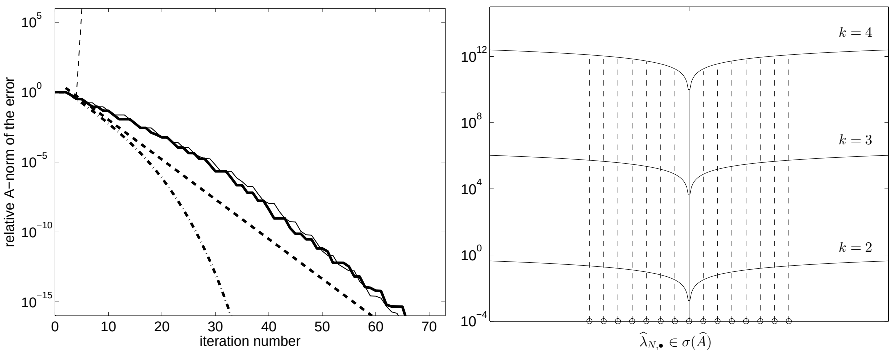

Fig. 5 Left: Whereas the exact CG convergence behaviour corresponding to $\widehat{A}, \widehat{b}$ (solid line) is both quantitatively and qualitatively different from the exact CG convergence behaviour corresponding to A, b (dash-dotted line), it nicely matches the finite precision CG computation (bold solid line) using A, b. The composite polynomial bound (29) (bold dashed line) is irrelevant and the bound (39) (dashed line) becomes after several iterations meaningless due to huge values of the composite polynomial in the neighborhood of the outlying eigenvalues of A. Right: Using the logarithmic vertical scale we plot a detail of the absolute values of the composite polynomial (with restriction to the values in the interval $[10^{-4}, 10^{13}]$) corresponding to the k-th iteration with $k = 2, 3$ and 4. The values of the composite polynomial at the eigenvalues $\widehat{\lambda}_{N,s}$, $s = 1, \ldots, l$ clustered around the largest eigenvalue λ_N blow up even for the smallest degrees of the corresponding shifted Chebyshev polynomial $\chi_{k-m}(\lambda)$ ($k - m = 1$ and 2). Here the width of the cluster around λ_N is $4\varepsilon \, \|A\| \approx 10^{-9}$.

$N = 40$ with $m = 2$ large outlying eigenvalues $\lambda_{N-1} = 10^4$, $\lambda_N = 10^6$ and with the eigenvalues $\lambda_1, \ldots, \lambda_{N-2}$ determined using

$$\lambda_i = \lambda_1 + \frac{i-1}{N-m-1} \left(\lambda_{N-m} - \lambda_1\right) \rho_{in}^{N-m-i} \quad i = 2, \ldots, N-m-1 \tag{40}$$

with $\rho_{in} = 0.9$, $\lambda_1 = 0.1$ and $\lambda_{N-2} = 1$. We compare it with exact CG convergence behaviour (solid line) corresponding to the associated vector $\widehat{b}$ and matrix $\widehat{A}$ with $\varDelta = 2\varepsilon \, \|A\|$ and $l = 15$, where $\varepsilon = 2^{-52}$ is machine roundoff unit; cf. [29, p. 126]. In agreement with [29] we observe quantitative and qualitative similarity of both convergence curves. The composite polynomial bound (29) (bold dashed line) with $m = 2$ (i.e. considering 2 largest eigenvalues of the matrix A as outliers) is for the finite precision computations irrelevant and the associated bound (39) (dashed line) practically immediately blows up. The latter is a consequence of the evaluation of the composite polynomial at the eigenvalues of $\widehat{A}$ clustered around the outlying eigenvalues of A as visualized in the right part of the figure.

The spectral upper bound applicable to finite precision CG computations based on the minimization problem (36) was investigated, following [27, 29], by Notay in [46]. He considered the composite polynomial where the part dealing with the outlying eigenvalues has possibly many roots in the neighborhood of the large outlying eigenvalues. The paper presents an estimate of the number of iterations needed to deal with the outlying eigenvalues as the number of iterations increases. This requires estimating the frequency of forming multiple copies of the large outlying eigenvalues, which unavoidably uses partially empirical arguments and requires knowledge of all large outlying eigenvalues. The paper [46] instructively demonstrates that *a-priori* investigation of the CG rate of convergence, which aims at realistic results including effects of rounding errors, is inevitably rather technical. Consequently, a practical application of a realistic *a-priori* analysis which is not specialized to some particular cases is limited.

5 Other shortcomings of composite polynomial bounds

In this section we will comment and numerically demonstrate several other drawbacks of the composite polynomial bound (29). Our observations can be summarized in the following points.

a) The composite polynomial bound (29) by construction does not depend on distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. In contrast to that, a finite precision CG behaviour can significantly depend on this distribution.
b) Unlike the bound (29), finite precision CG computations depend on the position of the large outlying eigenvalues.
c) The failure of the composite polynomial bound (29) in finite precision CG computations can occur even for a small size and/or conditioning of the problem.

In the numerical illustrations below we used diagonal matrices A and the right hand side b of all ones.

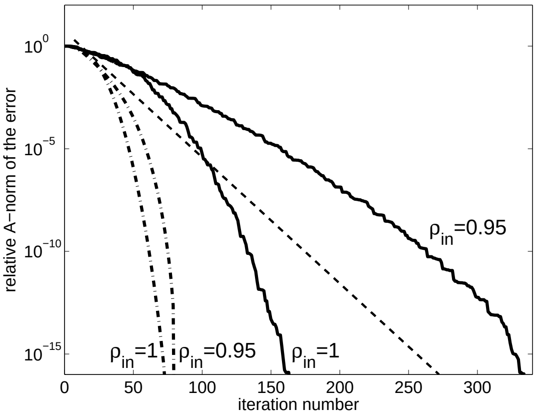

Fig. 6 Unlike the composite polynomial bound (dashed line), both exact (dash-dotted lines) and finite precision (bold solid lines) CG convergence behaviour are sensitive to the change of distribution of the eigenvalues in the interval $[\lambda_1, \lambda_{N-m}]$. In finite precision computations the difference between the uniform distribution with $\rho_{in} = 1$ and the distribution with $\rho_{in} = 0.95$ is significant.

Point a) In Figure 6 we compare CG computations applied to two problems with the same outlying eigenvalues, the same effective condition number $\kappa_m(A) = \lambda_{N-m}/\lambda_1$ but with different distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. Computations were performed using diagonal matrices of dimension $N = 80$ with $m = 7$ large outlying eigenvalues $\lambda_{N-6}, \ldots, \lambda_N$ and the eigenvalue λ_{N-7} determined using (34) with $\lambda_1 = 0.1,\ \lambda_N = 10^5$ and $\rho \equiv \rho_{out} = 0.3$. The eigenvalues $\lambda_2, \ldots, \lambda_{N-8}$ are distributed in the interval $[\lambda_1, \lambda_{N-7}]$ either uniformly or using (40) with $\rho_{in} = 0.95$.

The composite polynomial bound (29) with $m = 7$ (dashed line) is the same for both computations, as it does not reflect the distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. On the contrary, the convergence of the CG method depends in exact arithmetic slightly (dash-dotted lines) and in finite precision arithmetic very significantly (bold solid lines) on the distribution of *all* eigenvalues, including those in the interval $[\lambda_1, \lambda_{N-m}]$.

Point b) As mentioned in the previous paragraph, the convergence behaviour of the CG method depends on distribution of all eigenvalues. Thus the position of the outlying eigenvalues is of importance. In Figure 7 we plot the finite precision CG convergence curves (bold solid lines) and CG behaviour assuming exact arithmetic (dash-dotted lines) using the diagonal matrices of dimension $N = 50$ whose largest eigenvalue $\lambda_N = 10$ respectively $\lambda_N = 10^8$ is considered as *the only outlier* and the eigenvalues $\lambda_1, \ldots, \lambda_{N-1}$ are distributed uniformly within the interval $[\lambda_1, \lambda_{N-1}]$, $\lambda_1 = 0.1,\ \lambda_{N-1} = 0.3$.

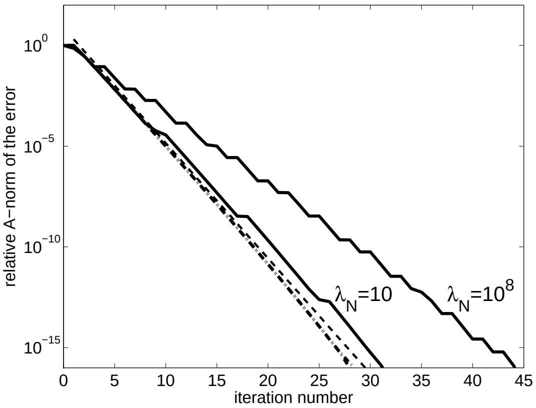

Fig. 7 Finite precision CG computations (bold solid lines) are, in contrast to the exact CG convergence behaviour (dash-dotted lines), sensitive to the position of the single large outlying eigenvalue λ_N. The frequency of forming multiple approximations of the largest eigenvalue is seriously affected by its position. The bounds based on the composite polynomial (28) (dashed line) can fail also in the presence of only a single large outlier.

The exact CG convergence behaviour is in both cases nearly identical. The delay of convergence in the finite precision CG computation with the outlying eigenvalue $\lambda_N = 10^8$ is naturally more significant than with the outlying eigenvalue $\lambda_N = 10$. This happens due to more frequent occurrence of the multiple approximations of the largest eigenvalue. Thus the information about the *number* of eigenvalues lying above some given number $\bar{\lambda}$ (as used, e.g., in [53, Corollary 2.2] or [38, p. 4]) is without further analysis of the problem not sufficient for estimating the actual convergence rate of finite precision CG computations. A single large outlying eigenvalue *can* affect the "asymptotic" rate of convergence. The composite polynomial bound (29) can fail even in this case.

Point c) Depending on the distribution of eigenvalues, the composite convergence bound can fail even for small and well-conditioned problems. We will use diagonal matrices with spectrum determined in the following way. We consider 4 different problems with $N = 30$ or 100 and $\lambda_N = 10$ or 10^6. The $m = 8$ large outlying eigenvalues $\lambda_{N-7}, \ldots, \lambda_N$ and the eigenvalue λ_{N-8} are given by (34) with $\lambda_1 = 0.1$, $\rho_{out} = 0.6$ for $N = 30$, $\rho_{out} = 0.2$ for $N = 100$. The rest of the eigenvalues is distributed in the interval $[\lambda_1, \lambda_{N-8}]$ using (40) with $\rho_{in} = 0.8$.

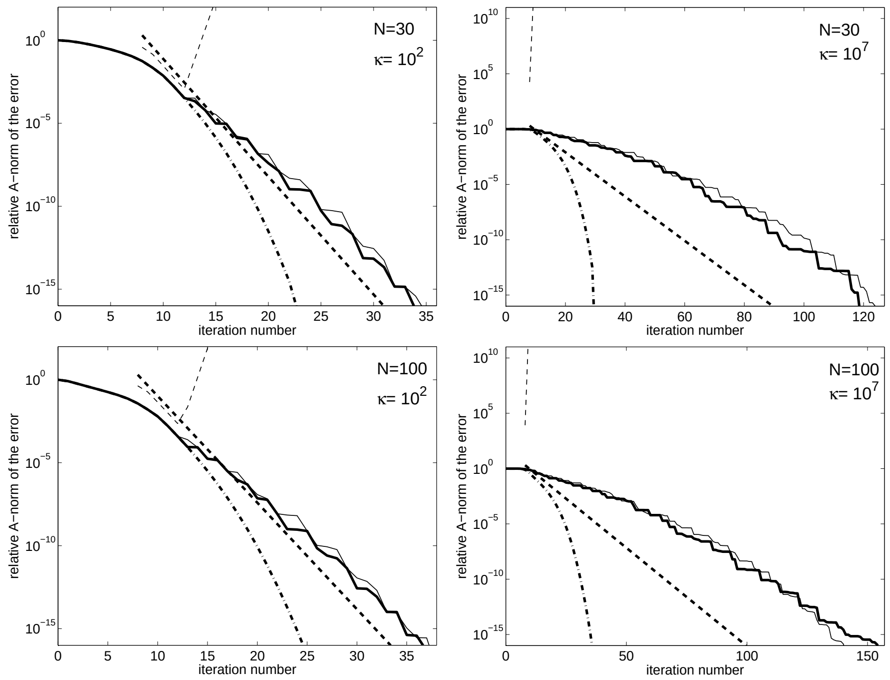

Fig. 8 The failure of the composite polynomial bound (bold dashed line) in finite precision CG computations (bold solid line) for well-conditioned (left) resp. ill-conditioned (right) smaller (top) and larger (bottom) problems. The exact CG convergence behaviour corresponding to $\widehat{A}$ (solid line) matches the finite precision CG computations performed using A and it differs both qualitatively and quantitatively from the exact CG convergence behaviour corresponding to A (dash-dotted line). The upper bound (39) (dashed line) which evaluates the composite polynomial in the neighborhood of outliers gives no relevant information.

Each of the subplots in Figure 8 shows that the composite polynomial bound (bold dashed line) and the finite precision CG convergence behaviour (bold solid line) have a little in common. We also plot the exact CG convergence behaviour (solid line) corresponding to the matrix $\widehat{A}$ which is determined using $\varDelta = \varepsilon \, \|A\|$ and $l = 15$. Similarly as in Section 4 we observe that it qualitatively matches the finite precision CG computations. The associated upper bound (39) (dashed line) becomes after several iterations meaningless.

6 Concluding remarks

This paper demonstrates that the composite polynomial bound (29) based on a Chebyshev polynomial and a fixed part having roots at large outlying eigenvalues of A has, in general, a little in common with actual finite precision CG computations. Related to that, CG method applied to a problem $Ax = b$ with a spectrum of the matrix A consisting of t tiny clusters does not necessarily produce a good approximation to the solution x within t steps. Many more steps may be needed, depending on the position of the individual clusters (this holds in exact arithmetic as well as in finite precision arithmetic). Our experimental illustration use small examples with diagonal matrices. In our opinion this makes the message appealing also for computations with real data.

Although this paper concentrates on bounds based on Chebyshev polynomials, the main point that the large outlying eigenvalues can challenge the relevance of *a-priori* CG convergence rate analysis when applied to practical computations is valid in general. Any *a-priori* CG convergence rate analysis is based on a substantial simplification of the very complex phenomena. We must admit this fact and verify any conclusion drawn from such analysis by justification of the assumptions incorporated in the whole development. *A-priori* convergence bounds are often used in connection with evaluation of preconditioning strategies and their optimality. Here the validity of the bounds in the presence of rounding errors and the *tightness* of the bounds should be taken as a strict requirement, otherwise the conclusions are not mathematically justified. There is an obvious exception, when preconditioning ensures very fast convergence, so that the tightness of the bounds does not matter. In such cases rounding errors have no chance to spoil significantly the computation.

In order to limit the effects of rounding errors, it would be useful to avoid pro-actively presence of large outlying eigenvalues in the spectrum of the preconditioned matrix; cf. [8]. Reorthogonalization procedures known from the Lanczos method for computing several dominating eigenvalues are in the CG context not generally applicable for efficiency reasons. They might be worth investigating, however, together with combined arithmetic techniques, in parallel implementations.

Finally, actual error in CG computations should be estimated and analyzed *a-posteriori*. This field has been thoroughly investigated by Golub and his collaborators, with early works [17,18]; see also [13–15]. As pointed out in [55], important steps in this direction can be found already in the original paper by Hestenes and Stiefel [33]. As in the *a-priori* analysis, the *a-posteriori* estimates and bounds can not be reliably applied to practical computations unless they are accompanied by a thorough rounding error analysis; see the arguments and examples given in [25,55,42]. For a survey we refer, e.g., to [41, Sections 3.3 and 5.3], [24, Chapter 12]. In the context of numerical solution of partial differential equations, the *a-posteriori* analysis of the algebraic iterations should be incorporated into the *a-posteriori* analysis of the whole solution process; see, e.g. the recent survey [1] and some possible challenges related to applications of CG formulated in [37, Chapter 5].

As in numerical solution of partial differential equations, *a-priori* and *a-posteriori* analysis has its place also in the iterative algebraic computations. In both fields *reliability* is the key requirement.

Acknowledgments. The authors are grateful to Jörg Liesen, Jan Papež and anonymous referee for useful comments.

References

1. Arioli, M., Liesen, J., Miedlar, A., and Strakoš, Z. Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems. *GAMM Mitt. Ges. Angew. Math. Mech.* (to appear).
2. Axelsson, O. A class of iterative methods for finite element equations. *Comput. Methods Appl. Mech. Engrg. 9* (1976), 123–127.
3. Axelsson, O. *Iterative solution methods.* Cambridge University Press, Cambridge, 1994.
4. Axelsson, O. A generalized conjugate gradient minimum residual method with variable preconditioners. In *Advanced mathematics: computations and applications (Novosibirsk, 1995).* NCC Publ., Novosibirsk, 1995, pp. 14–25.
5. Axelsson, O. Optimal preconditioners based on rate of convergence estimates for the conjugate gradient method. *Numer. Funct. Anal. Optim. 22* (2001), 277–302.
6. Axelsson, O., and Kaporin, I. On the sublinear and superlinear rate of convergence of conjugate gradient methods. *Numer. Algorithms 25* (2000), 1–22.
7. Axelsson, O., and Karátson, J. Equivalent operator preconditioning for elliptic problems. *Numer. Algorithms 50* (2009), 297–380.
8. Axelsson, O., and Lindskog, G. On the eigenvalue distribution of a class of preconditioning methods. *Numer. Math. 48* (1986), 479–498.
9. Beckermann, B., and Kuijlaars, A. B. J. On the sharpness of an asymptotic error estimate for conjugate gradients. *BIT 41* (2001), 856–867.
10. Beckermann, B., and Kuijlaars, A. B. J. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal. 39* (2001), 300–329.
11. Beckermann, B., and Kuijlaars, A. B. J. Superlinear CG convergence for special right-hand sides. *Electron. Trans. Numer. Anal. 14* (2002), 1–19.
12. Brezinski, C. *Projection Methods for Systems of Equations*, vol. 7 of *Studies in Computational Mathematics*. North-Holland, Amsterdam, 1997.
13. Brezinski, C. Error estimates for the solution of linear systems. *SIAM J. Sci. Comput. 21* (1999), 764–781.
14. Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. Computable error bounds and estimates for the conjugate gradient method. *Numer. Algorithms 25* (2000), 75–88.
15. Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. An iterative method with error estimators. *J. Comput. Appl. Math. 127* (2001), 93–119.
16. Dahlquist, G., and Björck, Å. *Numerical methods in scientific computing. Vol. I.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2008.

17. Dahlquist, G., Eisenstat, S. C., and Golub, G. H. Bounds for the error of linear systems of equations using the theory of moments. *J. Math. Anal. Appl. 37* (1972), 151–166.
18. Dahlquist, G., Golub, G. H., and Nash, S. G. Bounds for the error in linear systems. In *Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978)*, vol. 15 of *Lecture Notes in Control and Information Sci.* Springer, Berlin, 1979, pp. 154–172.
19. Daniel, J. W. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal. 4* (1967), 10–26.
20. Deuflhard, P. Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, vol. 180 of *Contemp. Math.* American Mathematical Society, Providence, RI, 1994, pp. 29–42.
21. Engeli, M., Ginsburg, T., Rutishauser, H., and Stiefel, E. Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. *Mitt. Inst. Angew. Math. Zürich. No. 8* (1959), 107.
22. Faber, V., Manteuffel, T. A., and Parter, S. V. On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations. *Adv. in Appl. Math. 11* (1990), 109–163.
23. Flanders, D. A., and Shortley, G. Numerical determination of fundamental modes. *J. Appl. Phys. 21* (1950), 1326–1332.
24. Golub, G. H., and Meurant, G. *Matrices, Moments and Quadrature with Applications*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, 2010.
25. Golub, G. H., and Strakoš, Z. Estimates in quadratic formulas. *Numer. Algorithms 8* (1994), 241–268.
26. Gratton, S., Titley-Peloquin, D., Toint, P., and Tshimanga, J. Linearizing the method of conjugate gradients. Technical Report naXys-15-2012, Namur Centre for Complex Systems, FUNDP–University of Namur, Belgium (2012).
27. Greenbaum, A. Behaviour of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl. 113* (1989), 7–63.
28. Greenbaum, A. *Iterative Methods for Solving Linear Systems*, vol. 17 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1997.
29. Greenbaum, A., and Strakoš, Z. Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM J. Matrix Anal. Appl. 13* (1992), 121–137.
30. Günnel, A., Herzog, R., and Sachs, E. A Note on Preconditioners and Scalar Products for Krylov Methods in Hilbert Space. (preprint).
31. Gutknecht, M. H., and Strakoš, Z. Accuracy of two three-term and three two-term recurrences for Krylov space solvers. *SIAM J. Matrix Anal. Appl. 22* (2000), 213–229.
32. Hackbusch, W. *Iterative Solution of Large Sparse Systems of Equations*, vol. 95 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994. Translated and revised from the 1991 German original.
33. Hestenes, M. R., and Stiefel, E. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards 49* (1952), 409–436.
34. Hiptmair, R. Operator preconditioning. *Comput. Math. Appl. 52* (2006), 699–706.
35. Jennings, A. Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Math. Appl. 20* (1977), 61–72.
36. Lanczos, C. Chebyshev polynomials in the solution of large-scale linear systems. In *Proceedings of the Association for Computing Machinery, Toronto, 1952* (1953), Sauls Lithograph Co. (for the Association for Computing Machinery), Washington, D. C., pp. 124–133.
37. Liesen, J., and Strakoš, Z. *Krylov subspace methods: principles and analysis.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2012.
38. Mardal, K.-A., and Winther, R. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl. 18* (2011), 1–40.
39. Markoff, A. Démonstration de certaines inégalités de M. Tchébychef. *Math. Ann. 24* (1884), 172–180.
40. Meurant, G. *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations.* vol. 19 of Software, Environments, and Tools. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2006.

41. Meurant, G., and Strakoš, Z. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer. 15* (2006), 471–542.
42. Meurant, G., and Tichý, P. On computing quadrature-based bounds for the A-norm of the error in conjugate gradients. *Numer. Algorithms 62* (2013), 163–191.
43. Naiman, A. E., Babuška, I. M., and Elman, H. C. A note on conjugate gradient convergence. *Numer. Math. 76* (1997), 209–230.
44. Naiman, A. E., and Engelberg, S. A note on conjugate gradient convergence. II, III. *Numer. Math. 85* (2000), 665–683, 685–696.
45. Nevanlinna, O. *Convergence of iterations for linear equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1993.
46. Notay, Y. On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math. 65* (1993), 301–317.
47. O'Leary, D. P., Strakoš, Z., and Tichý, P. On sensitivity of Gauss-Christoffel quadrature. *Numer. Math. 107* (2007), 147–174.
48. Paige, C. C. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Math. Appl. 18* (1976), 341–349.
49. Paige, C. C. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra and Its Applications 34* (1980), 235–258.
50. Richardson, L. F. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Phil. Trans. Roy. Soc. London A, 210* (1911), 307–357.
51. Rivlin, T. J. *Chebyshev Polynomials*, second ed. Pure and Applied Mathematics. John Wiley & Sons, New York, 1990.
52. Silvester, D. J., and Simoncini, V. An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. *ACM Trans. Math. Software 37* (2011), Art. 42, 22.
53. Spielman, D. A., and Woo, J. A note on preconditioning by low-stretch spanning trees. *Computing Research Repository* (2009).
54. Strakoš, Z. On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl. 154/156* (1991), 535–549.
55. Strakoš, Z., and Tichý, P. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal. 13* (2002), 56–80.
56. Strakoš, Z., and Tichý, P. Error estimation in preconditioned conjugate gradients. *BIT 45* (2005), 789–817.
57. Tyrtyshnikov, E. E. *A Brief Introduction to Numerical Analysis*. Birkhäuser, Boston, 1997.
58. van der Sluis, A., and van der Vorst, H. A. The rate of convergence of conjugate gradients. *Numer. Math. 48* (1986), 543–560.
59. van der Vorst, H. A. Iterative solution methods for certain sparse linear systems with a nonsymmetric matrix arising from PDE-problems. *J. Comput. Phys. 44* (1981), 1–19.
60. Varga, R. S. *Matrix iterative analysis*, expanded ed., vol. 27 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2000.
61. Vorobyev, Y. V. *Methods of Moments in Applied Mathematics*. Translated from the Russian by Bernard Seckler. Gordon and Breach Science Publishers, New York, 1965.
62. Winther, R. Some superlinear convergence results for the conjugate gradient method. *SIAM J. Numer. Anal. 17* (1980), 14–17.
63. Young, D. On Richardson's method for solving linear systems with positive definite matrices. *J. Math. Physics 32* (1954), 243–255.