SIAM J. OPTIM.
Vol. 31, No. 1, pp. 887–914




# THE CONVERGENCE OF THE GENERALIZED LANCZOS TRUST-REGION METHOD FOR THE TRUST-REGION SUBPROBLEM*

ZHONGXIAO JIA† AND FA WANG‡

**Abstract.** Solving the trust-region subproblem (TRS) plays a key role in numerical optimization and many other applications. The generalized Lanczos trust-region (GLTR) method is a well-known Lanczos type approach for solving a large-scale TRS. The method projects the original large-scale TRS onto a sequence of lower dimensional Krylov subspaces, whose orthonormal bases are generated by the symmetric Lanczos process, and computes approximate solutions from the underlying subspaces. There have been some a priori bounds available for the errors of the approximate solutions and approximate objective values obtained by the GLTR method, but no a priori bound exists on the errors of the approximate Lagrangian multipliers and the residual norms of approximate solutions obtained by the GLTR method. In this paper, a general convergence theory of the GLTR method is established for the TRS in the easy case, showing that the a priori bounds for these four quantities are closely interrelated and the one for the *computable* residual norm is of crucial importance in both theory and practice as it can predict the sizes of other three *uncomputable* errors reliably. Numerical experiments demonstrate that our bounds are realistic and predict the convergence rates of the four quantities accurately.

**Key words.** trust-region subproblem, GLTR method, a priori bound, easy case, hard case, Chebyshev polynomial, eigenvalue problem, symmetric Lanczos process

**AMS subject classifications.** 90C20, 90C30, 65K05, 65F10

**DOI.** 10.1137/19M1279691

## 1. Introduction.

Consider the solution of the trust-region subproblem (TRS)

$$\min_{\|s\|_B \le \Delta} q(s) \quad \text{with} \quad q(s) = g^T s + \frac{1}{2} s^T A s, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric, the nonzero $g \in \mathbb{R}^n$, $\Delta > 0$ is the trust-region radius, $B \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and the norm $\|s\|_B = \|B^{1/2}s\| = \sqrt{s^T B s}$. For $B = I$, the identity matrix, $\|\cdot\|_B = \|\cdot\|$ is the usual 2-norm. Problem (1.1) arises from nonlinear numerical optimization [3, 22], where $q(s)$ is a quadratic model of $\min f(s)$ at the current approximate solution, $A$ is Hessian, and $g$ is the gradient of $f$ at the current approximate solution, and many others, e.g., general-form Tikhonov regularization of ill-posed problems [25, 26], graph partitioning problems [15], the constrained eigenvalue problem [11], and the Levenberg–Marquardt algorithm for solving nonlinear least squares problems [22]. The matrix $B$ is often constructed to impose a smoothness condition on a solution to (1.1) for the ill-posed problem and to incorporate scaling of variables in optimization [25]. For instance, it is argued in [3] that a good choice is $B = J^{-T}J^{-1}$ for some invertible matrix $J$ or the Hermitian polar factor of $A$ [16].

*Received by the editors August 6, 2019; accepted for publication (in revised form) January 5, 2021; published electronically March 16, 2021.
https://doi.org/10.1137/19M1279691
**Funding:** This work was supported in part by the National Natural Science Foundation of China, grant 11771249.
†Corresponding author. Department of Mathematical Sciences, Tsinghua University, 100084 Beijing, China (jiazx@tsinghua.edu.cn).
‡Department of Mathematical Sciences, Tsinghua University, 100084 Beijing, China (wangfa15@mails.tsinghua.edu.cn).

The following results [3, 21] provide a theoretical basis for many TRS solution algorithms and give necessary and sufficient conditions, called the optimal conditions, for the solution of TRS (1.1).

THEOREM 1.1. *A vector $s_{opt}$ is a solution to* (1.1) *if and only if there exists the optimal Lagrangian multiplier $\lambda_{opt} \geq 0$ such that*

$$\|s_{opt}\|_B \leqslant \Delta, \tag{1.2}$$

$$(A + \lambda_{opt} B)s_{opt} = -g, \tag{1.3}$$

$$\lambda_{opt}(\Delta - \|s_{opt}\|_B) = 0, \tag{1.4}$$

$$A + \lambda_{opt} B \succeq 0, \tag{1.5}$$

*where the notation $\succeq 0$ indicates that a symmetric matrix is semipositive definite.*

TRS algorithms for solving (1.1) have been extensively studied for a few decades and can be classified as the following four categories, in which most of the algorithms in the first three categories are mentioned in [1].

- *Accurate methods for dense problems.* The Moré–Sorensen method [21] iteratively solves symmetric positive definite linear systems by the Cholesky factorizations. It is highly efficient and accurate for small to medium sized dense problems.
- *Accurate methods for large problems.* Algorithms in [25, 26, 28] iteratively compute the smallest eigenvalue of the matrix $\left(\begin{smallmatrix} \alpha & g^T \\ g & A \end{smallmatrix}\right)$, where $\alpha$ is an adjusted parameter. Another approach due to [24] solves TRS via semidefinite programming, and a modification of the Moré–Sorensen method using Taylor series is presented in [10]. The generalized Lanczos trust-region (GLTR) method [8] solves a large-scale TRS by a Lanczos type approach. Other accurate methods include subspace projection methods; see, e.g., [6, 14]. The GLTR method may be the most popular and efficient one among them.
- *Approximate methods for large problems.* Steihaug [29] and Toint [31] independently propose a truncated conjugate gradient (TCG) method, and Yuan [32] proves that the function reduction obtained at the point produced by this method is at least half of that obtained at the function minimizer when $A$ is symmetric positive definite. If $A$ is symmetric indefinite, an approximate solution must reach the trust-region boundary, and this makes TCG a globally but slowly convergent algorithm [22].
- *Promising eigenvalue-based methods.* The method due to Gander, Golub, and von Matt [11] reduces TRS (1.1) with $B = I$ to a single quadratic eigenvalue problem, which is linearized to a standard eigenvalue problem of size $2n$. Using a different derivation, Adachi et al. [1] extend the eigenvalue-based formulation in [11] to TRS (1.1) with $B \neq I$ possibly and formulate it as a generalized eigenvalue problem of size $2n$. A solution to (1.1) can be found by the rightmost eigenvalue and the associated eigenvector of the $2n \times 2n$ matrix pair.

Notice that the general TRS (1.1) with $B \neq I$ is mathematically equivalent to a standard TRS (1.1) with $B = I$ through the substitutions

$$\widehat{A} \leftarrow B^{-\frac{1}{2}} A B^{-\frac{1}{2}}, \quad \widehat{g} \leftarrow B^{-\frac{1}{2}} g, \quad \text{and} \quad c \leftarrow B^{\frac{1}{2}} s, \tag{1.6}$$

which transform the general TRS (1.1) with $B \neq I$ into the standard TRS with $B = I$:

$$\min_{\|c\| \leq \Delta} \widehat{q}(c) \quad \text{with} \quad \widehat{q}(c) = \widehat{g}^T c + \frac{1}{2} c^T \widehat{A} c. \tag{1.7}$$

For ease of presentation, we will first consider TRS (1.1) with $B = I$ when analyzing the convergence of the GLTR method. Then based on the equivalence relationship between (1.1) and (1.7), we shall extend all the convergence results on the GLTR method with $B = I$ to the case that $B \neq I$.

The GLTR method and other projection methods have been shown to be efficient for a large-scale TRS; see, e.g., [2, 5, 8]. Let $s_{opt}$ be a solution to TRS (1.1) and $s_k$ be the approximate solution from the underlying $k+1$ dimensional Krylov subspace $\mathcal{K}_k(g, A) = \text{span}\{g, Ag, \ldots, A^k g\}$ obtained by the GLTR method. By Theorem 1.1, there is an optimal Lagrangian multiplier $\lambda_k$ for each reduced TRS problem of dimension $k+1$ (cf. (3.7)) onto $\mathcal{K}_k(g, A)$. There are four central convergence problems: how fast the three errors $|\lambda_{opt} - \lambda_k|$, $\|s_k - s_{opt}\|$, $q(s_k) - q(s_{opt})$ and the residual norm $\|(A + \lambda_k I)s_k + g\|$ of the approximate solution $\lambda_k, s_k$ of (1.3) decrease as $k$ increases. For $\|s_k - s_{opt}\|$ and $q(s_k) - q(s_{opt})$, some a priori bounds have been derived in [33]. However, there have been no a priori bounds for $|\lambda_{opt} - \lambda_k|$ and $\|(A + \lambda_k I)s_k + g\|$ hitherto. The only known result on $\lambda_k$ is that $\lambda_k$ increases monotonically with $k$ and is bounded from above by $\lambda_{opt}$ [19]. Therefore, we always have $|\lambda_{opt} - \lambda_k| = \lambda_{opt} - \lambda_k \geq 0$. Notice that the errors $\lambda_{opt} - \lambda_k$, $\|s_k - s_{opt}\|$ and $q(s_k) - q(s_{opt})$ are *uncomputable* in practice. Therefore, the *computable* residual norm $\|(A + \lambda_k I)s_k + g\|$ is of particular importance in both theory and practice as its size is commonly used to measure the convergence of the GLTR method. We mention that a mixed bound is given for $\lambda_{opt} - \lambda_k$ in [34, Lemma 3.4]. However, it is easy to check that the mixed bound in [34] does not exhibit any decreasing tendency and even can never be small unless the symmetric Lanczos process breaks down, in which case the bound is trivially zero.

Strikingly, it has recently been shown that, under the mild conditions that $\|s_k\| = \|s_{opt}\| = \Delta$, the solution of (1.1) in the easy case is mathematically equivalent to computing the rightmost eigenpair of a certain matrix eigenvalue problem of size $2n$ [1]. Among others, such mathematical equivalence makes us realize that, at iteration $k$, the GLTR method amounts to solving a certain matrix eigenvalue problem of size $2(k+1)$ by projecting the $2n \times 2n$ matrix eigenvalue problem onto a special $2(k+1)$ dimensional subspace in $\mathbb{R}^{2n}$ constructed by $\mathcal{K}_k(g, A)$ used in the GLTR method. At iteration $k$, unlike the GLTR method, one can simultaneously obtain the optimal $\lambda_k$ and the solution $s_k$ to the projected TRS. This key observation is our starting point to study the convergence of the GLTR method. A note is that we are mainly concerned with $\sin\angle(s_k, s_{opt})$ rather than $\|s_k - s_{opt}\|$. The sine is a standard measure when considering the error of an eigenvector and its approximations in the context of the matrix eigenvalue problem [30]. As will be clear later, if $\|s_k\| = \|s_{opt}\| = \Delta$, then we have $\sin\angle(s_k, s_{opt}) \approx \|s_k - s_{opt}\|/\Delta$ once they start to become fairly small, that is, these two errors are essentially equal. For $\|s_k\| = \|s_{opt}\| = \Delta$, the authors of [1] measure the error of $s_k$ and $s_{opt}$ by the sine of angle $\angle(s_k, s_{opt})$ in their experiments.

The main contributions in this paper are the establishment of the two a priori bounds for $\lambda_{opt} - \lambda_k$, that of an a priori bound for $\|(A + \lambda_k I)s_k + g\|$ and that of a bound for $\sin\angle(s_k, s_{opt})$. As by-products and for our use in deriving an a priori bound for $\|(A + \lambda_k I)s_k + g\|$, we also derive new sharp bounds for $\|s_k - s_{opt}\|$ and $q(s_k) - q(s_{opt})$. The first a priori bound for $\lambda_{opt} - \lambda_k$ is the background for establishing the second much sharper one. When establishing the first a priori bound for $\lambda_{opt} - \lambda_k$ and the a priori bound for $\sin\angle(s_k, s_{opt})$, we need to solve the best uniform polynomial approximation problem to the rational function $\frac{1}{(x-\eta)^2}$ with $x \in [-1, 1]$ and $\eta > 1$. We will make use of a generating function of $\frac{1}{(x-\eta)^2}$ with Chebyshev polynomials of

the second kind [4] and find a quasi-optimal approximation polynomial in order to obtain an accurate estimate for the best approximation error.

Our results will show that the bounds for $\|(A+\lambda_k I)s_k+g\|$ and $\|s_k-s_{opt}\|$ decrease exactly at the same rate, and the bounds for $\lambda_{opt}-\lambda_k$ and $\|q(s_k)-q(s_{opt})\|$ tend to zero also at the same rate but their convergence rates are the squares of the former two. Precisely, let $\|(A+\lambda_k I)s_k+g\| = tol$. Then $\|s_k-s_{opt}\| = \mathcal{O}(tol)$, $q(s_k)-q(s_{opt}) = \mathcal{O}(tol^2)$, and $\lambda_{opt}-\lambda_k = \mathcal{O}(tol^2)$ with the constants in the $\mathcal{O}(\cdot)$ being modest. Numerical results will demonstrate that our a priori bounds not only predict the convergence rates of the three errors and residual norms but also estimate their values accurately. Notice that the *computable* $\|(A+\lambda_k I)s_k+g\|$ is used as a stopping criterion for the GLTR method [8, 9]. The a priori bound for $\|(A+\lambda_k I)s_k+g\|$ is of crucial importance in both theory and practice since its size can predict the sizes of the other three *uncomputable* errors reliably. An important implication of these results is that if one is concerned with the approximate size(s) of uncomputable $\lambda_{opt}-\lambda_k$, $\|s_k-s_{opt}\|$ and $q(s_k)-q(s_{opt})$ or one of them, then a proper stopping criterion *tol* can be prescribed for $\|(A+\lambda_k I)s_k+g\|$.

In section 2, we give some preliminaries and introduce the equivalence of the solution of (1.1) and a certain $2n\times 2n$ matrix eigenvalue problem. We review the GLTR method in section 3. Section 4 is devoted to a priori bounds for $\lambda_{opt}-\lambda_k$. A priori bounds for $\sin\angle(s_k,s_{opt})$ and $\|(A+\lambda_k I)s_k+g\|$ are presented in section 5. In section 6, we extend the convergence results on the GLTR method for solving the standard TRS (1.7) to the GLTR method for solving the general TRS (1.1). In section 7, we report numerical experiments to confirm that our bounds estimate not only the convergence rates but also the three errors and residual norms obtained by the GLTR method accurately. Finally, we conclude the paper in section 8.

Throughout this paper, denote by the superscript $T$ the transpose of a matrix or vector, by $\|\cdot\|$ the 2-norm of a matrix or vector, by $I$ the identity matrix with order clear from the context, and by $e_i$ the $i$th column of $I$.

## 2. Preliminaries.

**2.1. A solution to TRS (1.1).** Suppose that $A = V\Lambda V^T$ is the eigendecomposition of $A$, where $V=(v_1,v_2,\ldots,v_n)$ is orthogonal and $\Lambda=\mathrm{diag}(\alpha_1,\alpha_2,\ldots,\alpha_n)$ with the $\alpha_i$ being the eigenvalues of $A$ labeled as $\alpha_1\ge\alpha_2\ge\cdots\ge\alpha_n$ and $v_i$ the eigenvector of $A$ associated with $\alpha_i$.

If $A+\lambda_{opt}I\succ 0$, the solution $s_{opt}$ to TRS (1.1) is unique and $s_{opt}=-(A+\lambda_{opt}I)^{-1}g$. Particularly, if $A$ is positive definite and $\lambda_{opt}=0$, then $s_{opt}=-A^{-1}g$ with $\|s_{opt}\|\le\Delta$. These correspond to the "easy case" [3, 8, 21, 22] or "nondegenerate case" [14]. If $A$ is symmetric semipositive definite or indefinite and

$$g\perp\mathcal{N}(A-\alpha_n I),$$

the null space of $A-\alpha_n I$, i.e., the eigenspace of $A$ associated with the smallest eigenvalue $\alpha_n$, then we have the following definition [3, 8, 22].

Definition 2.1 (hard case). *TRS* (1.1) *is in the hard case if* $g\perp\mathcal{N}(A-\alpha_n I)$, *and the optimal Lagrangian multiplier* $\lambda_{opt}=-\alpha_n$.

The hard case is also called the "degenerate case" [14]. In this case, one must have $\alpha_n\le 0$, which means that $A$ must be semipositive definite when $\alpha_n=0$ and indefinite when $\alpha_n<0$, and TRS (1.1) may have multiple optimal solutions [22, pp. 87–88], which must reach the trust-region boundary and can be characterized as

$$s_{opt} = -(A - \alpha_n I)^{\dagger} g + \xi v_n, \tag{2.1}$$

where $v_n \in \mathcal{N}(A - \alpha_n I)$ and $\|v_n\| = 1$, $\|(A - \alpha_n I)^{\dagger} g\| < \Delta$, and the superscript † denotes the Moore–Penrose generalized inverse. From $\|s_{opt}\| = \Delta$ and the orthogonality of two terms in the right-hand side of (2.1), it is known that the scalar $\eta$ satisfies

$$\xi^2 = \Delta^2 - \|(A - \alpha_n I)^{\dagger} g\|^2 > 0. \tag{2.2}$$

As we have seen, in the hard case, we not only need to compute the eigenspace of $A$ associated with the smallest eigenvalue $\alpha_n$ but also need to solve the singular symmetric semi–positive definite system $(A - \alpha_n I)s = -g$ for its minimum 2-norm solution $-(A - \alpha_n I)^{\dagger} g$. We then form $s_{opt}$ in (2.1) by determining $\xi$ via (2.2). The hard case has been studied for years; see, e.g., [1, 7, 8, 21, 22, 24].

Clearly, as far as the solution of TRS (1.1) is concerned, the easy case and the hard case are mathematically and numerically different, and they have to be treated separately.

**2.2. The equivalence of the TRS and a matrix eigenvalue problem.** When the solution $s_{opt}$ to the general TRS (1.1) in the easy case reaches the trust-region boundary, Adachi et al. [1] have proven that TRS (1.1) is mathematically equivalent to a certain generalized eigenvalue problem of order $2n$ when $B \neq I$, which, for $B = I$, reduces to the standard eigenvalue problem of the augmented matrix

$$M = \begin{pmatrix} -A & \frac{gg^T}{\Delta^2} \\ I & -A \end{pmatrix} \in \mathbb{R}^{2n \times 2n}. \tag{2.3}$$

Let $\mu_1, \mu_2, \ldots, \mu_{2n}$ be the eigenvalues of $M$ labeled as

$$\mathrm{Re}(\mu_1) \geq \mathrm{Re}(\mu_2) \geq \cdots \geq \mathrm{Re}(\mu_{2n}), \tag{2.4}$$

where $\mathrm{Re}(\cdot)$ is the real part of a scalar. The following result in [1] establishes a key relationship between the TRS solution $s_{opt}$ and a specific eigenpair of $M$.

THEOREM 2.2 ([1]). *Let $(\lambda_{opt}, s_{opt})$ satisfy Theorem 1.1 with $\|s_{opt}\| = \Delta$. Then the rightmost eigenvalue $\mu_1$ of $M$ is real and simple, and $\mu_1 = \lambda_{opt}$. Let $y = (y_1^T, y_2^T)^T$ be the corresponding unit length eigenvector of $M$ with $y_1, y_2 \in \mathbb{R}^n$, and suppose that $g^T y_2 \neq 0$. Then the unique TRS solution is*

$$s_{opt} = -\frac{\Delta^2}{g^T y_2} y_1. \tag{2.5}$$

*Remark* 2.1. Adachi et al. [1] have proven that $g^T y_2 = 0$ corresponds to the hard case, i.e., $\lambda_{opt} = -\alpha_n$ and $g \perp \mathcal{N}(A - \alpha_n I)$. Therefore, in the easy case, $g^T y_2 \neq 0$ is guaranteed, so that (2.5) is well defined.

**3. The generalized Lanczos trust-region method [8].** For (1.1) large, the GLTR method iteratively solves a sequence of smaller projected TRS's

$$\min_{s \in \mathcal{S}_k, \|s\| \leq \Delta} q(s) \quad \text{with} \quad q(s) = g^T s + \frac{1}{2} g^T A g, \tag{3.1}$$

where the $k+1$ dimensional subspace $\mathcal{S}_k$ is chosen as the Krylov subspace,

$$\mathcal{S}_k = \mathcal{K}_k(A, g) \doteq \mathrm{span}\{g, Ag, A^2 g, \ldots, A^k g\}, \tag{3.2}$$

generated by $g$ and $A$, and we use the solution $s_k$ of TRS (3.1) to approximate the solution $s_{opt}$ of (1.1).

The GLTR method consists of two phases. In the first phase, it starts with the TCG method with the zero initial guess [29, 31]. When $A$ is positive definite and $\|A^{-1}g\| \leq \Delta$, which corresponds to $\lambda_{opt} = 0$, the method ultimately returns a *converged* approximate solution $s_k$ to $s_{opt} = -A^{-1}g$. In this case, the trust-region constraint is inactive and $\|s_k\| < \Delta$ monotonically increases with $k$. Therefore, the standard conjugate gradient (CG) method solves (1.1), and its convergence theory is directly applicable. If the TCG iterate exceeds the trust-region boundary or a negative curvature is encountered, the trust-region constraint is activated, meaning that $A$ is semipositive definite or indefinite and $\lambda_{opt} > 0$. The GLTR method then enters the second phase and switches to the Lanczos method that accurately solves (3.1). It proceeds in such a way until $s_k$ converges to $s_{opt}$.

In what follows, we always assume that the TCG method does not solve (3.1) exactly, i.e., $\lambda_{opt} > 0$, and one must use the Lanczos method to solve (3.1) accurately from the first iteration upward.

At iteration $k$, the GLTR method exploits the symmetric Lanczos process to generate an orthonormal basis $\{q_i\}_{i=0}^k$ of $\mathcal{S}_k$ defined by (3.2), which can be written in matrix form

$$AQ_k = Q_k T_k + \beta_{k+1} q_{k+1} e_{k+1}^T, \tag{3.3}$$

$$Q_k^T g = \beta_0 e_1, \ \beta_0 = \|g\|, \ g = \beta_0 q_0, \tag{3.4}$$

where $Q_k = (q_0, q_1, \ldots, q_k)$ is orthonormal and the matrix

$$T_k = Q_k^T A Q_k = \begin{pmatrix} \delta_0 & \beta_1 & & & \\ \beta_1 & \delta_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \delta_{k-1} & \beta_k \\ & & & \beta_k & \delta_k \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}. \tag{3.5}$$

We shall consider vectors of form

$$s = Q_k h \in \mathcal{S}_k. \tag{3.6}$$

Let $s_k = Q_k h_k$ solve the projected TRS (3.1). It then follows from (3.6) and (3.3)–(3.5) that $h_k$ solves the *reduced* TRS

$$\min_{\|h\|\leq\Delta} \phi(h) \ \text{ with } \ \phi(h) = \beta_0 e_1^T h + \frac{1}{2} h^T T_k h \tag{3.7}$$

and

$$q(s_k) = \phi(h_k). \tag{3.8}$$

From Theorem 1.1, the vector $h_k$ is a solution to (3.7) if and only if there exists the optimal Lagrangian multiplier $\lambda_k \geq 0$ such that

$$\|h_k\| \leqslant \Delta, \tag{3.9}$$

$$(T_k + \lambda_k I) h_k = -\beta_0 e_1, \tag{3.10}$$

$$\lambda_k(\Delta - \|h_k\|) = 0, \tag{3.11}$$

$$T_k + \lambda_k I \succeq 0. \tag{3.12}$$

Suppose that TRS (1.1) is in the easy case. Then Gould et al. [8] have proven that TRS (3.7) is always in the easy case provided that the symmetric Lanczos process does not break down. Furthermore, they have shown that the residual norm of $\lambda_k$ and $s_k$ as approximate solutions of (1.3) for $B = I$ satisfies

$$\|(A + \lambda_k I)s_k + g\| = \beta_{k+1}|e_{k+1}^T h_k| \tag{3.13}$$

and that if the symmetric Lanczos process breaks down at iteration $k_{\max}$ for the first time, then $s_{k_{\max}} = s_{opt}$ and $\lambda_{k_{\max}} = \lambda_{opt}$; see Theorem 5.7 in [8].

Before proceeding, we should elaborate more on the GLTR method when (1.1) is in the hard case. From Definition 2.1 of the hard case and the expression (2.1) of $s_{opt}$, it is known that the subspace $\mathcal{S}_k$ defined by (3.2) does not contain any information on $\mathcal{N}(A - \alpha_n I)$, that is, $Q_k$ generated by the Lanczos process starting with $g$ cannot extract any information on $\mathcal{N}(A - \alpha_n I)$. Therefore, $s_k = Q_k h_k$ does not contain any component of $\mathcal{N}(A - \alpha_n I)$, so that $s_k$ *never converges* to $s_{opt}$ in the hard case as it is only an approximation to $-(A - \alpha_n I)^\dagger g$ by noting that $\|(A - \alpha_n I)^\dagger g\| < \Delta$ and $\xi \neq 0$ in (2.1). We comment that $-(A - \alpha_n I)^\dagger g$ is the minimum 2-norm solution to $(A - \alpha_n I)s = -g$. In other words, the GLTR method itself cannot solve TRS in the hard case. Gould et al. [8] have proposed a *restarting* strategy of the GLTR method to cure this problem theoretically, which is, however, unlikely to be of practical interest as they have addressed; see [8, Theorem 5.8] and the discussions that follow.

In the next two sections, just as done in [33], we shall consider the convergence of the GLTR method in the easy case, establish a priori bounds for $\lambda_{opt} - \lambda_k$, $\sin\angle(s_k, s_{opt})$, and $\|(A + \lambda_k I)s_k + g\|$, and prove how they decrease as $k$ increases.

**4. A priori bounds for $\lambda_{opt} - \lambda_k$.** We establish a priori bounds for $\lambda_{opt} - \lambda_k$ in this section. Hitherto the only known result on $\lambda_k$ is qualitative and states that $\lambda_k$ increases monotonically with $k$ and is bounded from above by $\lambda_{opt}$ [19], i.e.,

$$0 \leq \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{k_{\max}} = \lambda_{opt},$$

where $k_{\max}$ is the first iteration at which the symmetric Lanczos process breaks down.

We point out that when $\lambda_{opt} > 0$, the situation $\lambda_k = 0$ might happen in the early stage of the GLTR method but one must have $\lambda_k > 0$ and $s_k$ thus reaches the trust-region boundary as $k$ increases. As a matter of fact, $\lambda_k = 0$ implies that $s_k$ is identical to the Steihaug's TCG iterate with the zero initial guess and is in the interior of the trust-region [29]. In such case, it is shown in [22, pp. 71–72, pp. 171–172] and [29] that $s_1$ is a Cauchy point, that is, it is parallel to $-g$, $\|s_k\|$ increases, and $q(s_k) \leq q(s_1)$ decreases monotonically with $k$; moreover, TCG must terminate at a point $s_k$ for which $q(s_k) \leq q(s_k^c)$ with $s_k^c$ a Cauchy point, that is, the reduction in the objective function $q(s)$ equals or exceeds that of the Cauchy point, meaning that TCG is globally convergent and converges at least as fast as the steepest descent method.

Nevertheless, as has been addressed in [8], considerable experience has shown that $\lambda_k > 0$ frequently happens during the first few iterations, and often the first. The experiments in [33] have indicated that this happens at $k = 1$ without exception, and our experiments have also confirmed this. In fact, for $A$ indefinite, this phenomenon can be theoretically justified: $T_k$ must be indefinite soon as its largest and smallest eigenvalues, two Ritz values of $A$ with respect to $\mathcal{S}_k$, approximate the *positive* largest and *negative* smallest ones of $A$, respectively, as $k$ increases [20, 23], so that, by (3.12), we must have $\lambda_k > 0$ soon; moreover, it is very likely that $\lambda_1 > 0$ since the two

eigenvalues of $T_1$ approximate the positive largest and negative smallest eigenvalues of $A$, respectively. Therefore, without loss of generality we always assume that $\|s_k\| = \|h_k\| = \Delta$ and $\lambda_k > 0$ when considering the convergence of the GLTR method.

Define the matrix

$$\widetilde{Q}_k = \begin{pmatrix} Q_k & \\ & Q_k \end{pmatrix} \tag{4.1}$$

with $Q_k$ defined by (3.3). Obviously, $\widetilde{Q}_k$ is column orthonormal, and its columns span the $2(k+1)$ dimensional subspace denoted as

$$\widetilde{\mathcal{S}}_k = \begin{pmatrix} \mathcal{S}_k & 0 \\ 0 & \mathcal{S}_k \end{pmatrix} \subset \mathbb{R}^{2n}. \tag{4.2}$$

Let the $2(k+1) \times 2(k+1)$ matrix

$$M_k = \widetilde{Q}_k^T M \widetilde{Q}_k \tag{4.3}$$

with $M$ defined by (2.3). Then it is straightforward that

$$M_k = \begin{pmatrix} -T_k & \frac{\beta_0^2 e_1 e_1^T}{\Delta^2} \\ I & -T_k \end{pmatrix} \tag{4.4}$$

with $T_k$ defined by (3.5) and $\beta_0 = \|g\|$. Therefore, $M_k$ is the orthogonal projection matrix of $M$ onto $\widetilde{\mathcal{S}}_k$ in the orthonormal basis $\{(q_i^T, 0)^T\}_{i=0}^k$ and $\{(0, q_i^T)^T\}_{i=0}^k$.

Let $\mu_i^{(k)}$, $i = 1, 2, \ldots, 2(k+1)$, be the eigenvalues of $M_k$ labeled as

$$\mathrm{Re}(\mu_1^{(k)}) \geq \mathrm{Re}(\mu_2^{(k)}) \geq \cdots \geq \mathrm{Re}(\mu_{2(k+1)}^{(k)}).$$

By applying Theorem 2.2 to the reduced TRS (3.7) and the equivalent eigenvalue problem of the matrix $M_k$, it is known that $\mu_1^{(k)}$ is real simple and

$$\mu_1^{(k)} = \lambda_k. \tag{4.5}$$

Let $z^{(k)} = ((z_1^{(k)})^T, (z_2^{(k)})^T)^T$ with $z_1^{(k)}, z_2^{(k)} \in \mathbb{R}^{k+1}$ be the unit length eigenvector of $M_k$ associated with its rightmost eigenvalue $\mu_1^{(k)}$. Then

$$y^{(k)} = \widetilde{Q}_k z^{(k)} = \left((Q_k z_1^{(k)})^T, (Q_k z_2^{(k)})^T\right)^T = \left((y_1^{(k)})^T, (y_2^{(k)})^T\right)^T \tag{4.6}$$

is a Ritz vector of $A$ from the subspace $\widetilde{\mathcal{S}}_k$ and approximates the unit length eigenvector $y = (y_1^T, y_2^T)^T$ of $M$ associated with its rightmost real eigenvalue $\mu_1 = \lambda_{opt}$.

From structure (4.4) of $M_k$ and the definition of $z^{(k)}$, it is easy to show that $((z_2^{(k)})^T, (z_1^{(k)})^T)^T$ is the left eigenvector of $M_k$ corresponding to the real simple eigenvalue $\mu_1^{(k)} = \lambda_k$. From the definition of $z_k$ it is straightforward to verify that

$$z_2^{(k)} = (T_k + \lambda_k I)^{-1} z_1^{(k)}. \tag{4.7}$$

Therefore, by definition (cf. [30, p. 186]), the spectral condition number of $\mu_1^{(k)}$ is

$$\kappa(\lambda_k) = \frac{1}{2|(z_2^{(k)})^T z_1^{(k)}|} = \frac{1}{2(z_1^{(k)})^T (T_k + \lambda_k I)^{-1} z_1^{(k)}}. \tag{4.8}$$

Similarly, by structure (2.3) of $M$, it is easily justified that the vector $(y_2^T, y_1^T)^T$ is the left eigenvector of $M$ associated with the rightmost eigenvalue $\mu_1 = \lambda_{opt}$ and $y_2 = (A + \lambda_{opt} I)^{-1} y_1$. As a result, the spectral condition number of $\mu_1$ is

$$\kappa(\lambda_{opt}) = \frac{1}{2|y_2^T y_1|} = \frac{1}{2 y_1^T (A + \lambda_{opt} I)^{-1} y_1}. \tag{4.9}$$

By Theorem 2.2 and (4.6), the unique solution $s_k$ to TRS (3.1) is

$$s_k = Q_k h_k = Q_k \left( -\frac{\Delta^2}{(\beta_0 e_1)^T z_2^{(k)}} z_1^{(k)} \right) = -\frac{\Delta^2}{(\beta_0 e_1)^T z_2^{(k)}} y_1^{(k)}. \tag{4.10}$$

Denote by $\angle(u, \mathcal{S}_k)$ the acute angle between a nonzero vector $u$ and $\mathcal{S}_k$. Then

$$\sin \angle(u, \mathcal{S}_k) = \frac{\|(I - \pi_k) u\|}{\|u\|}, \tag{4.11}$$

where $\pi_k$ is the orthogonal projector onto $\mathcal{S}_k$ and $\|(I - \pi_k)u\|$ is the distance between $u$ and $\mathcal{S}_k$. In terms of Theorem 2.2 and (4.5), we have

$$\lambda_{opt} - \lambda_k = \mu_1 - \mu_1^{(k)}. \tag{4.12}$$

Let $\widetilde{\pi}_k = \widetilde{Q}_k \widetilde{Q}_k^T$ be the orthogonal projector onto $\widetilde{\mathcal{S}}_k$. Then $\widetilde{\pi}_k M \widetilde{\pi}_k$ is the restriction of $M$ to the subspace $\widetilde{\mathcal{S}}_k$ and its matrix representation is $M_k$ in the orthonormal basis $\{(q_i^T, 0)^T\}_{i=0}^k$ and $\{(0, q_i^T)^T\}_{i=0}^k$. The eigenvalues of $\widetilde{\pi}_k M \widetilde{\pi}_k$ restricted to $\widetilde{\mathcal{S}}_k$ are the eigenvalues of $M_k$, called the Ritz values, and the eigenvectors are the Ritz vectors of $M$ from $\widetilde{\mathcal{S}}_k$; see [27]. Therefore, a direct application of Theorem 3.8 in [17] to our context gives the following result.

LEMMA 4.1. *Let $\mu_1^{(k)} = \lambda_k$ and $\mu_1 = \lambda_{opt}$ be the rightmost eigenvalues of $M_k$ and $M$, respectively, and $y$ is the unit length eigenvector of $M$ with $\mu_1$, and suppose that $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. Then for $\sin \angle(y, \widetilde{\mathcal{S}}_k)$ small it holds that*

$$\lambda_{opt} - \lambda_k \le \kappa(\lambda_k) \widetilde{\gamma}_k \sin \angle(y, \widetilde{\mathcal{S}}_k) + \mathcal{O}(\sin^2 \angle(y, \widetilde{\mathcal{S}}_k)), \tag{4.13}$$

*where $\kappa(\lambda_k)$ is defined by (4.8), $\widetilde{\gamma}_k = \|\widetilde{\pi}_k M (I - \widetilde{\pi}_k)\|$.*

From (4.6) and (4.8), due to the orthonormality of $Q_k$, we obtain

$$\kappa(\lambda_k) = \frac{1}{2|(y_2^{(k)})^T y_1^{(k)}|},$$

which converges to $\kappa(\lambda_{opt})$ defined by (4.9) when $y^{(k)} \to y$. This is indeed the case, as will be shown in the next section. In the meantime, $\widetilde{\gamma}_k \le \|M\|$, a fixed constant. As a result, by this lemma, the convergence problem of $\lambda_k$ to $\lambda_{opt}$ becomes the problem of analyzing how $\sin \angle(y, \widetilde{\mathcal{S}}_k)$ decreases as $k$ increases.

Notice that

$$\sin^2 \angle(y, \widetilde{\mathcal{S}}_k) = \left\| (I - \widetilde{\pi}_k) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\|^2 = \|(I - \pi_k) y_1\|^2 + \|(I - \pi_k) y_2\|^2. \tag{4.14}$$

Therefore, in order to bound $\lambda_{opt} - \lambda_k$ and to show how it converges to zero as $k$ increases, we need to analyze both $\|(I - \pi_k) y_1\|$ and $\|(I - \pi_k) y_2\|$.

We first consider $\|(I - \pi_k) y_1\|$. In what follows, we denote by $\bar{P}_k$ the set of polynomials of degree not exceeding $k + 1$. We present the following result.

LEMMA 4.2. *The distance* $\|(I-\pi_k)s_{opt}\|$ *between* $s_{opt}$ *and* $\mathcal{S}_k = \mathcal{K}_k(A,g)$ *satisfies*

$$\|(I-\pi_k)s_{opt}\| = \min_{p_k\in\bar{P}_k, p_k(0)=1} \|p_k(A+\lambda_{opt}I)s_{opt}\| \tag{4.15}$$

*and*

$$\|(I-\pi_k)s_{opt}\| \le \|s_{opt}\|\epsilon_1^{(k)},\ k=0,1,\ldots,k_{\max}, \tag{4.16}$$

*where*

$$\epsilon_1^{(k)} = \min_{p_k\in\bar{P}_k, p_k(0)=1} \max_{1\le i\le n} |p_k(\alpha_i+\lambda_{opt})| \tag{4.17}$$

*with* $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_n$ *being the eigenvalues of* $A$. *Moreover,*

$$\epsilon_1^{(k)} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+1}, \tag{4.18}$$

*where* $\kappa = \frac{\alpha_1+\lambda_{opt}}{\alpha_n+\lambda_{opt}}$ *is the condition number of* $A+\lambda_{opt}I$.

*Proof.* Theorem 1.1 has shown that $s_{opt}$ satisfies the linear system $(A+\lambda_{opt})s_{opt} = -g$. Therefore, exploiting the shift invariance $\mathcal{K}_k(A,g) = \mathcal{K}_k(A+\lambda_{opt}I,g)$ and the eigendecomposition $A = V\Lambda V^T$ with $V$ being orthogonal, we obtain

$$\begin{aligned}
\|(I-\pi_k)s_{opt}\| &= \min_{s\in\mathcal{K}_k(A+\lambda_{opt}I,g)} \|s_{opt}-s\| \\
&= \min_{q\in\bar{P}_{k-1}} \|s_{opt}-q(A+\lambda_{opt}I)g\| \\
&= \min_{q\in\bar{P}_{k-1}} \|s_{opt}+q(A+\lambda_{opt}I)(A+\lambda_{opt})s_{opt}\| \\
&= \min_{p_k\in\bar{P}_k, p_k(0)=1} \|p_k(A+\lambda_{opt}I)s_{opt}\| \\
&\le \|s_{opt}\| \min_{p_k\in\bar{P}_k, p_k(0)=1} \|p_k(\Lambda+\lambda_{opt}I)\| = \|s_{opt}\|\epsilon_1^{(k)}
\end{aligned}$$

with the polynomial $p_k(\lambda) = 1+\lambda q(\lambda) \in \bar{P}_k$ and $p_k(0) = 1$.

Note that $A+\lambda_{opt}I$ is symmetric positive definite. Applying a standard estimate (cf. the book [12, Theorem 3.1.1, p. 51] to $\epsilon_1^{(k)}$, we obtain (4.18). ∎

Relation (2.5) shows that $y_1$ is the same as $s_{opt}$ up to a scaling. Therefore, replacing $s_{opt}$ in (4.15) and (4.16) by $y_1$ and exploiting (4.18), we have established the following bound for $\|(I-\pi_k)y_1\|$.

THEOREM 4.3. *Let* $y = (y_1^T, y_2^T)^T$ *be the unit length eigenvector of* $M$ *associated with its rightmost eigenvalue* $\mu_1$ *and let* $\kappa$ *be defined as in Lemma* 4.2. *Then*

$$\|(I-\pi_k)y_1\| \le 2\|y_1\|\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+1},\ k=0,1,\ldots,k_{\max}. \tag{4.19}$$

As it will turn out, an effective estimation of $\|(I-\pi_k)y_2\|$ is much more involved.

THEOREM 4.4. *With the previous notation, let* $\alpha_1$ *and* $\alpha_n$ *be the largest and smallest eigenvalues of* $A$. *Then*

$$\|(I-\pi_k)y_2\| \le \frac{4(\alpha_1+\lambda_{opt})}{(\alpha_1-\alpha_n)^2}\|y_1\|\epsilon_2^{(k)}, k=0,1,\ldots,k_{\max}, \tag{4.20}$$

*with*

$$\epsilon_2^{(k)} = \min_{q\in\bar{P}_{k-1}} \max_{x\in[-1,1]} \left| \frac{1}{(x-\eta)^2} - q(x) \right|, \tag{4.21}$$

$$\eta = \frac{\alpha_1 + \alpha_n + 2\lambda_{opt}}{\alpha_1 - \alpha_n} = \frac{\kappa+1}{\kappa-1} > 1. \tag{4.22}$$

*Proof.* Recall that $A = V\Lambda V^T$ with $V$ being orthogonal and $\Lambda = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$. From $(A + \lambda_{opt} I)s_{opt} = -g$ and (2.5), we obtain

$$\frac{\Delta^2}{g^T y_2}(A + \lambda_{opt} I)y_1 = g.$$

From (2.3) and the definition of $y$, we have

$$y_2 = (A + \lambda_{opt} I)^{-1} y_1,$$

which, together with $\mathcal{K}_k(A, g) = \mathcal{K}_k(A + \lambda_{opt} I, g)$ and the orthogonality of $V$, yields

$$\begin{aligned}
\|(I - \pi_k)y_2\| &= \min_{z\in\mathcal{K}_k(A+\lambda_{opt}I,g)} \|y_2 - z\| \\
&= \min_{q\in\bar{P}_{k-1}} \|y_2 - q(A + \lambda_{opt} I)g\| \\
&= \min_{q\in\bar{P}_{k-1}} \|(A + \lambda_{opt} I)^{-1} y_1 - \frac{\Delta^2}{g^T y_2}(A + \lambda_{opt} I)q(A + \lambda_{opt} I)y_1\| \\
&= \min_{p\in\bar{P}_{k-1}} \|(A + \lambda_{opt} I)[(A + \lambda_{opt} I)^{-2} - p(A + \lambda_{opt} I)]y_1\| \\
&\leq \|A + \lambda_{opt} I\| \min_{p\in\bar{P}_{k-1}} \|[(A + \lambda_{opt} I)^{-2} - p(A + \lambda_{opt} I)]y_1\| \\
&= \|A + \lambda_{opt} I\| \min_{p\in\bar{P}_{k-1}} \|V[(\Lambda + \lambda_{opt} I)^{-2} - p(\Lambda + \lambda_{opt} I)]V^T y_1\| \\
&\leq (\alpha_1 + \lambda_{opt})\|y_1\| \min_{p\in\bar{P}_{k-1}} \max_{z\in[\alpha_n,\alpha_1]} \left| \frac{1}{(z + \lambda_{opt})^2} - p(z) \right|.
\end{aligned}$$

Consider the variable transformation

$$z = \frac{\alpha_1 - \alpha_n}{2} x + \frac{\alpha_n + \alpha_1}{2},$$

which maps $x \in [-1, 1]$ to $z \in [\alpha_n, \alpha_1]$ in one-to-one correspondence. Then

$$\begin{aligned}
\min_{p\in\bar{P}_{k-1}} \max_{z\in[\alpha_n,\alpha_1]} \left| \frac{1}{(z + \lambda_{opt})^2} - p(z) \right| &= \min_{p\in\bar{P}_{k-1}} \max_{x\in[-1,1]} \left| \frac{4}{(\alpha_1 - \alpha_n)^2(x - \eta)^2} - p(x) \right| \\
&= \frac{4}{(\alpha_1 - \alpha_n)^2} \min_{q\in\bar{P}_{k-1}} \max_{x\in[-1,1]} \left| \frac{1}{(x - \eta)^2} - q(x) \right| \\
&= \frac{4}{(\alpha_1 - \alpha_n)^2} \epsilon_2^{(k)}. \qquad \square
\end{aligned} \tag{4.23}$$

$\epsilon_2^{(k)}$ is the error of the best uniform polynomial approximation from $\bar{P}_{k-1}$ to the rational function $\frac{1}{(x-\eta)^2}$ over the interval $[-1, 1]$ with $\eta > 1$. To our best knowledge, there seems no known explicit solution to such an approximation problem. Recall from (4.14) that $\sin\angle(y, \widetilde{\mathcal{S}}_k) > \|(I - \pi)y_1\|$. Therefore, it suffices to seek a quasi-optimal

polynomial approximation and prove that $\epsilon_2^{(k)}$ converges to zero as fast as bound (4.19) because this means that $\sin\angle(y,\widetilde{\mathcal{S}}_k)$ is of the same order as bound (4.19) for $\|(I-\pi)y_1\|$. To this end, exploiting Chebyshev polynomials of the second kind, we are able to establish a desired bound for $\epsilon_2^{(k)}$, which is as small as bound (4.19).

THEOREM 4.5. *Let* $t=\eta-\sqrt{\eta^2-1}$ *with* $\eta$ *as in* (4.22) *and* $\kappa=\frac{\alpha_1+\lambda_{opt}}{\alpha_n+\lambda_{opt}}$. *Then, for* $k=0,1,\ldots,k_{\max}$,

$$\epsilon_2^{(k)} \le \left(1+\frac{k+2}{|\ln t|}\right)\frac{4}{1-t^2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+3}, \tag{4.24}$$

$$\|(I-\pi_k)y_2\| \le \frac{16(\alpha_1+\lambda_{opt})\|y_1\|}{(\alpha_1-\alpha_n)^2(1-t^2)}\left(1+\frac{k+2}{|\ln t|}\right)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+3}. \tag{4.25}$$

*Proof.* For any $t\in(-1,1)$ and $x\in[-1,1]$ there is the following generating function [4, p. 215]:

$$\sum_{j=0}^{\infty}(j+1)t^jU_j(x)=\frac{1-t^2}{(1+t^2-2tx)^2}, \tag{4.26}$$

where $U_j(x)=\sin(j\arccos x)$ is the $j$th degree Chebyshev polynomial of the second kind [4, p. 212]. For $t=\eta-\sqrt{\eta^2-1}$, we have $1+t^2=2\eta t$, and (4.22) shows that

$$t=\eta-\sqrt{\eta^2-1}=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}<1. \tag{4.27}$$

Therefore, (4.26) becomes

$$\sum_{j=0}^{\infty}(j+1)t^jU_j(x)=\frac{1-t^2}{4t^2(x-\eta)^2}, \tag{4.28}$$

from which it follows that

$$\frac{1}{(x-\eta)^2}=\frac{4t^2}{1-t^2}\sum_{j=0}^{\infty}(j+1)t^jU_j(x).$$

Take the $k$th degree polynomial

$$p_k(x)=\frac{4t^2}{1-t^2}\sum_{j=0}^{k}(j+1)t^jU_j(x)\in\bar{P}_{k-1},$$

and $|U_j(x)|\le 1$ for $x\in[-1,1]$. Then from (4.21) we obtain

$$\begin{aligned}
\epsilon_2^{(k)} &\le \max_{x\in[-1,1]}\left|\frac{1}{(x-\eta)^2}-p_k(x)\right| = \max_{x\in[-1,1]}\left|\frac{4t^2}{1-t^2}\sum_{j=k+1}^{\infty}(j+1)t^jU_j(x)\right| \\
&\le \frac{4t^2}{1-t^2}\sum_{j=k+1}^{\infty}(j+1)t^j = \frac{4t^2}{1-t^2}\int_{k+1}^{\infty}(z+1)t^z\,dz \\
&= \frac{4t^2}{1-t^2}\left(\frac{z+1}{\ln t}t^z\Big|_{k+1}^{\infty}-t^z\Big|_{k+1}^{\infty}\right) \\
&= \left(1-\frac{k+2}{\ln t}\right)\frac{4t^{k+3}}{1-t^2}=\left(1+\frac{k+2}{|\ln t|}\right)\frac{4t^{k+3}}{1-t^2},
\end{aligned}$$

which, together with (4.20) and (4.23), leads to (4.24) and (4.25). ∎

Combining Lemma 4.1, (4.14), Theorem 4.3, and Theorem 4.5 ultimately leads to the following bounds.

THEOREM 4.6. *Suppose that* $\|s_{opt}\| = \|s_k\| = \Delta$. *Then*

$$\sin \angle(y, \widetilde{\mathcal{S}}_k) \le c_k \|y_1\| \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+1} \tag{4.29}$$

*and asymptotically*

$$\lambda_{opt} - \lambda_k \le c_k \kappa(\lambda_k) \widetilde{\gamma}_k \|y_1\| \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+1}, \tag{4.30}$$

*where*

$$c_k = 2 + \frac{16(\alpha_1 + \lambda_{opt})}{(\alpha_1 - \alpha_n)^2(1-t^2)} \left(1 + \frac{k+2}{|\ln t|}\right) \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2. \tag{4.31}$$

A priori bound (4.30) proves that $\lambda_{opt} - \lambda_k$ converges to zero as $k$ increases. Exploiting this bound, we can establish a much sharper a priori bound for $\lambda_{opt} - \lambda_k$. To this end, we first derive the following result.

THEOREM 4.7. *For* $k = 0, 1, \ldots, k_{max}$, *we have*

$$e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1 - e_1^T (T_k + \lambda_{opt} I)^{-1} e_1 \le \frac{4\Delta}{\beta_0} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)}. \tag{4.32}$$

*Proof.* Consider the symmetric positive definite linear system

$$(T_{k_{\max}} + \lambda_{opt} I) h = -\beta_0 e_1. \tag{4.33}$$

Taking $e_1$ as the starting vector, the $(k+1)$-step symmetric Lanczos process generates an orthonormal basis $\{e_i\}_{i=1}^{k+1}$ of the $(k+1)$-dimensional Krylov subspace

$$\mathcal{K}_{k+1}(T_{k_{\max}} + \lambda_{opt} I, e_1) = \text{span}\{e_1, (T_{k_{\max}} + \lambda_{opt} I) e_1, \ldots, (T_{k_{\max}} + \lambda_{opt} I)^k e_1\}$$

and the symmetric tridiagonal matrix $T_k + \lambda_{opt} I$. Define $E_k = (e_1, e_2, \ldots, e_{k+1})$. Then the orthogonal projection matrix of $T_{k_{\max}} + \lambda_{opt} I$ onto $\mathcal{K}_{k+1}(T_{k_{\max}} + \lambda_{opt} I, e_1)$ in the basis of $\{e_i\}_{i=1}^{k+1}$ is

$$T_k + \lambda_{opt} I = E_k^T (T_{k_{\max}} + \lambda_{opt} I) E_k.$$

Applying the symmetric Lanczos method [12, 20] to solving (4.33), for $k \le k_{\max}$ we obtain the projected problem

$$(T_k + \lambda_{opt} I) \tilde{y} = -\beta_0 e_1.$$

Write its solution as $\tilde{y}_k$. Then the symmetric Lanczos method computes $\tilde{h}_k = E_k \tilde{y}_k$ as an approximation to the solution $h_{k_{\max}}$ to (4.33).

Define the error $\varepsilon_k = h_{k_{\max}} - \tilde{h}_k$ and the residual $r_k = -\beta_0 e_1 - (T_{k_{\max}} + \lambda_{opt} I) \tilde{h}_k$ of (4.33). Note that the initial residual $r_0 = -\beta_0 e_1$. Then $\|r_0\| = \beta_0$ and

$$(T_{k_{\max}} + \lambda_{opt} I) \varepsilon_k = r_k.$$

From this relation and [20, Theorem 2.11] it follows that the square of $(T_{k_{\max}}+\lambda_{opt}I)$-norm error satisfies

$$\begin{aligned}\|\varepsilon_k\|^2_{(T_{k_{\max}}+\lambda_{opt}I)} &= \varepsilon_k^T(T_{k_{\max}}+\lambda_{opt}I)\varepsilon_k^T \\ &= r_k^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}r_k \\ &= \beta_0^2\left(e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1\right).\end{aligned}$$

As a result, we obtain

$$e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1 = \frac{\|\varepsilon_k\|^2_{(T_{k_{\max}}+\lambda_{opt}I)}}{\beta_0^2}. \tag{4.34}$$

Notice that the eigenvalues of $T_{k_{\max}}$ are the exact eigenvalues of $A$, which means that the smallest and largest eigenvalues of $T_{k_{\max}}+\lambda_{opt}I$ lie in $[\alpha_n+\lambda_{opt}, \alpha_1+\lambda_{opt}]$. Since the symmetric Lanczos method is mathematically equivalent to the CG method at the same iteration when the same initial guess on $h_{k_{\max}}$ is used, applying a standard estimate (cf. [12, Theorem 3.1.1] and [20, Theorem 2.30]) to $\|\varepsilon_k\|^2_{(T_{k_{\max}}+\lambda_{opt}I)}$ yields

$$\|\varepsilon_k\|^2_{(T_{k_{\max}}+\lambda_{opt}I)} \le 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)} \|\varepsilon_0\|^2_{(T_{k_{\max}}+\lambda_{opt}I)}.$$

Since $r_0 = -\beta_0 e_1$, the squared initial error

$$\|\varepsilon_0\|^2_{(T_{k_{\max}}+\lambda_{opt}I)} = r_0^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}r_0 = \beta_0^2 e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1.$$

Exploiting $\beta_0\|(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1\| = \|h_{k_{\max}}\| = \Delta$, we obtain

$$\beta_0^2 e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 \le \beta_0\|e_1\|\Delta = \beta_0\Delta.$$

Substituting the above three relations into (4.34) proves (4.32). ∎

THEOREM 4.8. *Suppose that* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then asymptotically*

$$\lambda_{opt} - \lambda_k \le \eta_{k1}\left(e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1\right) + \eta_{k2}\left(q(s_k) - q(s_{opt})\right) \tag{4.35}$$

*with*

$$\eta_{k1} = \frac{\beta_0^2}{\Delta^2+\beta_0^2 e_1^T(T_k+\lambda_{opt}I)^{-2}e_1} \le \frac{\beta_0^2(\alpha_1+\lambda_{opt})^2}{\beta_0^2+(\alpha_1+\lambda_{opt})^2\Delta^2}, \tag{4.36}$$

$$\eta_{k2} = \frac{2}{\Delta^2+\beta_0^2 e_1^T(T_k+\lambda_{opt}I)^{-2}e_1} \le \frac{2(\alpha_1+\lambda_{opt})^2}{\beta_0^2+(\alpha_1+\lambda_{opt})^2\Delta^2}. \tag{4.37}$$

*Proof.* From (3.10), we obtain

$$h_k = -\beta_0(T_k+\lambda_k I)^{-1}e_1$$

and $\|h_k\| = \beta_0\|(T_k+\lambda_k I)^{-1}e_1\| = \Delta$. By (3.7), we have $q(s_k) = \phi(h_k)$. Therefore,

$$
\begin{aligned}
q(s_k) &= -\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 + \frac{1}{2}\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} T_k (T_k + \lambda_k I)^{-1} e_1 \\
&= -\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 + \frac{1}{2}\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} (T_k + \lambda_k I - \lambda_k I)(T_k + \lambda_k I)^{-1} e_1 \\
&= -\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 + \frac{1}{2}\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 - \frac{1}{2}\lambda_k \beta_0^2 e_1^T (T_k + \lambda_k I)^{-2} e_1 \\
&= -\frac{1}{2}\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 - \frac{1}{2}\lambda_k \beta_0^2 e_1^T (T_k + \lambda_k I)^{-2} e_1 \\
&= -\frac{1}{2}\beta_0^2 e_1^T (T_k + \lambda_k I)^{-1} e_1 - \frac{1}{2}\lambda_k \Delta^2. \qquad (4.38)
\end{aligned}
$$

By assumption and (3.7), it is known that $\|h_{k_{max}}\| = \Delta$,

$$s_{k_{\max}} = Q_{k_{\max}} h_{k_{\max}} = s_{opt}, \quad \lambda_{k_{\max}} = \lambda_{opt}, \quad q(s_{k_{\max}}) = q(s_{opt}) = \phi(h_{k_{max}})$$

and the eigenvalues $T_{k_{\max}}$ are the exact eigenvalues of $A$. In this case, (4.38) is

$$q(s_{opt}) = -\frac{1}{2}\beta_0^2 e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1 - \frac{1}{2}\lambda_{opt}\Delta^2. \qquad (4.39)$$

Subtracting the two sides of (4.38) and (4.39) yields

$$(\lambda_{opt} - \lambda_k)\Delta^2 = \beta_0^2 \left( e_1^T (T_k + \lambda_k I)^{-1} e_1 - e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1 \right) + 2\left( q(s_k) - q(s_{opt}) \right). \qquad (4.40)$$

Since $\|(T_k + \lambda_{opt} I)^{-1}\| \leq \frac{1}{\alpha_n + \lambda_{opt}}$ and (4.30) has proven that $\lambda_{opt} - \lambda_k$ is nonnegative and tends to zero as $k$ increases, we must have $(\lambda_{opt} - \lambda_k)\|(T_k + \lambda_{opt} I)^{-1}\| < 1$, i.e., $\lambda_{opt} - \lambda_k \leq \alpha_n + \lambda_{opt}$, for $k$ sufficiently large. Moreover, since $\lambda_k \to \lambda_{opt}$, by a continuity argument, we have

$$e_1^T (T_k + \lambda_k I)^{-1} e_1 - e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1 \to e_1^T (T_k + \lambda_{opt} I)^{-1} e_1 - e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1,$$

where the quantity in the right-hand side has been shown by (4.34) to be strictly *negative* for all $k = 0, 1, \ldots, k_{\max} - 1$. Therefore, the first term $e_1^T (T_k + \lambda_k I)^{-1} e_1 - e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1$ in the right-hand side of (4.40) must become *nonpositive* as $k$ increases, so that the inequality

$$(\lambda_{opt} - \lambda_k)\Delta^2 \leq \beta_0^2 \left( e_1^T (T_{k_{\max}} + \lambda_{opt} I)^{-1} e_1 - e_1^T (T_k + \lambda_k I)^{-1} e_1 \right) + 2\left( q(s_k) - q(s_{opt}) \right) \qquad (4.41)$$

must hold for $k$ sufficiently large.

Next we analyze $e_1^T (T_k + \lambda_k I)^{-1} e_1$. Since $(\lambda_{opt} - \lambda_k)\|(T_k + \lambda_{opt} I)^{-1}\| < 1$ for $k$ sufficiently large, the series expansion of $\left( (I - (\lambda_{opt} - \lambda_k)(T_k + \lambda_{opt} I)^{-1}) \right)^{-1}$ gives

$$
\begin{aligned}
(T_k + \lambda_k I)^{-1} &= (T_k + \lambda_{opt} I + (\lambda_k - \lambda_{opt}) I)^{-1} \\
&= \left( (T_k + \lambda_{opt} I)(I - (\lambda_{opt} - \lambda_k)(T_k + \lambda_{opt} I)^{-1}) \right)^{-1} \\
&= \left( (I - (\lambda_{opt} - \lambda_k)(T_k + \lambda_{opt} I)^{-1}) \right)^{-1} (T_k + \lambda_{opt} I)^{-1} \\
&= \left( I + (\lambda_{opt} - \lambda_k)(T_k + \lambda_{opt} I)^{-1} + \mathcal{O}\left( (\lambda_{opt} - \lambda_k)^2 \right) \right) (T_k + \lambda_{opt} I)^{-1} \\
&= (T_k + \lambda_{opt} I)^{-1} + (\lambda_{opt} - \lambda_k)(T_k + \lambda_{opt} I)^{-2} + \mathcal{O}((\lambda_{opt} - \lambda_k)^2).
\end{aligned}
$$

Therefore, we have

$$e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_k I)^{-1}e_1 = e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1 - (\lambda_{opt}-\lambda_k)e_1^T(T_k+\lambda_{opt}I)^{-2}e_1 - \mathcal{O}((\lambda_{opt}-\lambda_k)^2),$$

which is *nonnegative* for $k$ sufficiently large. Substituting this relation into (4.41) and dropping the smaller term $\mathcal{O}((\lambda_{opt}-\lambda_k)^2)$ in the resulting left-hand side gives

$$\lambda_{opt}-\lambda_k \le \eta_{k1}\left(e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1\right) + \eta_{k2}\left(q(s_k)-q(s_{opt})\right)$$

with $\eta_{k1}$ and $\eta_{k2}$ defined by (4.36) and (4.37), respectively, which proves (4.35).

Since $T_k+\lambda_{opt}I$ is symmetric positive definite and its eigenvalues lie between $\alpha_n+\lambda_{opt}$ and $\alpha_1+\lambda_{opt}$, respectively, we have

$$\frac{1}{(\alpha_1+\lambda_{opt})^2} \le e_1^T(T_k+\lambda_{opt}I)^{-2}e_1 \le \frac{1}{(\alpha_n+\lambda_{opt})^2}.$$

As a result, from the forms of $\eta_{k1}$ and $\eta_{k2}$, it is straightforward to obtain

$$\eta_{k1} \le \frac{\beta_0^2(\alpha_1+\lambda_{opt})^2}{\beta_0^2+(\alpha_1+\lambda_{opt})^2\Delta^2}, \quad \eta_{k2} \le \frac{2(\alpha_1+\lambda_{opt})^2}{\beta_0^2+(\alpha_1+\lambda_{opt})^2\Delta^2}.$$ ∎

Relation (4.35) shows that bounding $\lambda_{opt}-\lambda_k$ amounts to bounding $e_1^T(T_{k_{\max}}+\lambda_{opt}I)^{-1}e_1 - e_1^T(T_k+\lambda_{opt}I)^{-1}e_1$ and $q(s_k)-q(s_{opt})$ separately. We have established the a priori bound (4.32) for the former one. Now we investigate $q(s_k)-q(s_{opt})$. Zhang, Shen, and Li [33, Theorem 4.3] have given the following result. Starting with it, we can derive an a priori bound for $q(s_k)-q(s_{opt})$.

LEMMA 4.9 ([33]). *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then*

$$0 \le q(s_k) - q(s_{opt}) \le 2(\alpha_1+\lambda_{opt})\|\tilde{s}-s_{opt}\|^2 \tag{4.42}$$

*for any nonzero* $\tilde{s} \in \mathcal{K}_k(A,g)$.

THEOREM 4.10. *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then*

$$0 \le q(s_k) - q(s_{opt}) \le 8(\alpha_1+\lambda_{opt})\Delta^2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)}. \tag{4.43}$$

*Proof.* Relation (4.42) has shown that

$$q(s_k) - q(s_{opt}) \le 2(\alpha_1+\lambda_{opt}) \min_{\tilde{s}\in\mathcal{K}_k(A,g)} \|\tilde{s}-s_{opt}\|^2. \tag{4.44}$$

By definition, we have

$$\min_{\tilde{s}\in\mathcal{K}_k(A,g)} \|\tilde{s}-s_{opt}\|^2 = \|(I-\pi_k)s_{opt}\|^2, \tag{4.45}$$

where $\pi_k$ is the orthogonal projector onto $\mathcal{K}_k(A,g)$. From the above relation and Lemma 4.2, it is immediate that

$$\min_{\tilde{s}\in\mathcal{K}_k(A,g)} \|\tilde{s}-s_{opt}\|^2 \le 4\|s_{opt}\|^2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)} = 4\Delta^2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)}.$$

Substituting it into (4.44) yields (4.43). ∎

Bound (4.43) is slightly worse and clearer than bound (4.18a) in [33].

Substituting bound (4.43) into (4.35) and bound (4.32) into (4.35) ultimately leads to the following a priori bound for $\lambda_{opt} - \lambda_k$.

THEOREM 4.11. *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then asymptotically*

$$\lambda_{opt} - \lambda_k \leq \left( \frac{4\eta_{k1}\Delta}{\beta_0} + 8(\alpha_1 + \lambda_{opt})\eta_{k2}\Delta^2 \right) \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2(k+1)} \tag{4.46}$$

*with the constants* $\eta_{k1}$ *and* $\eta_{k2}$ *defined by* (4.36) *and* (4.37), *respectively.*

This theorem indicates that, except for the modest constant, $\lambda_{opt} - \lambda_k$ converges to zero at least as fast as $(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2(k+1)}$. Therefore, bound (4.46) is much sharper than bound (4.30) and is approximately the square of the latter.

**5. A priori bounds for $\sin\angle(s_k, s_{opt})$ and $\|(A + \lambda_k I)s_k + g\|$.** Suppose that $\|s_{opt}\| = \|s_k\| = \Delta$. Then the measures $\sin\angle(s_k, s_{opt})$ and $\|s_k - s_{opt}\|/\Delta$ are essentially equivalent once they start to become fairly small since

$$\begin{aligned} \frac{\|s_k - s_{opt}\|^2}{\Delta^2} &= \frac{s_k^T s_k}{\Delta^2} + \frac{s_{opt}^T s_{opt}}{\Delta^2} - 2\frac{s_k^T s_{opt}}{\Delta^2} = 1 + 1 - 2\cos\angle(s_k, s_{opt}) \\ &= 4\sin^2\frac{\angle(s_k, s_{opt})}{2} \approx \sin^2\angle(s_k, s_{opt}). \end{aligned} \tag{5.1}$$

It is seen from (4.10) and (2.5) that $s_k$ and $s_{opt}$ are the same as $y_1^{(k)}$ and $y_1$ up to scaling, respectively. As a result, we have

$$\sin\angle(s_k, s_{opt}) = \sin\angle(y_1^{(k)}, y_1). \tag{5.2}$$

We take two steps to estimate $\sin\angle(s_k, s_{opt})$. First, we bound $\sin\angle(s_k, s_{opt})$ in terms of $\sin\angle(y^{(k)}, y)$ with $y^{(k)}$ defined by (4.6). Second, we establish an a priori bound for $\sin\angle(y^{(k)}, y)$, showing how it converges to zero as $k$ increases. To this end, we need the following result [13, Lemma 2.3].

LEMMA 5.1 ([13]). *Let* $u = \binom{u_1}{u_2}$ *and* $\tilde{u} = \binom{\tilde{u}_1}{\tilde{u}_2}$ *with* $u_i$, $\tilde{u}_i \in \mathbb{C}^n$ *for* $i = 1, 2$ *and* $\|u_1\| = \|\tilde{u}_1\| = 1$. *Then*

$$\sin\angle(u_1, \tilde{u}_1) \leq \min\{\|u\|, \|\tilde{u}\|\} \sin\angle(u, \tilde{u}).$$

THEOREM 5.2. *For the unit length eigenvector* $y = (y_1^T, y_2^T)^T$ *of* $M$ *associated with the rightmost eigenvalue* $\mu_1 = \lambda_{opt}$ *and the unit length vector* $y^{(k)}$ *defined by* (4.6), *we have*

$$\sin\angle(s_k, s_{opt}) \leq \frac{1}{\|y_1\|} \sin\angle(y^{(k)}, y). \tag{5.3}$$

*Proof.* From (5.2), since

$$\sin\angle(s_k, s_{opt}) = \sin\angle(y_1^{(k)}, y_1) = \sin\angle\left( \frac{y_1^{(k)}}{\|y_1^{(k)}\|}, \frac{y_1}{\|y_1\|} \right)$$

with the unit length vectors $y_1^{(k)}/\|y_1^{(k)}\|$ and $y_1/\|y_1\|$, by definition (4.6) of $y^{(k)}$ and Lemma 5.1 we obtain

$$\sin\angle(s_k, s_{opt}) = \sin\angle\left(\frac{y_1^{(k)}}{\|y_1^{(k)}\|}, \frac{y_1}{\|y_1\|}\right) \le \min\left\{\frac{1}{\|y_1\|}, \frac{1}{\|y_1^{(k)}\|}\right\} \sin\angle\left(\frac{y^{(k)}}{\|y_1^{(k)}\|}, \frac{y}{\|y_1\|}\right)$$

$$\le \frac{1}{\|y_1\|}\sin\angle(y^{(k)}, y). \tag{5.4}$$ ∎

Bound (5.3) indicates that the convergence of $\sin\angle(s_k, s_{opt})$ is dictated by that of $\sin\angle(y^{(k)}, y)$ as $k$ increases.

Recall that $(\mu_1, y)$ and $(\mu_1^{(k)}, z^{(k)})$ are real simple eigenpairs of $M$ and $M_k$, respectively, and $(\mu_1^{(k)}, y^{(k)})$ is the Ritz pair approximating the eigenpair. Let the columns of $Z_\perp^{(k)}$ be an orthonormal basis of the orthogonal complement of the subspace spanned by $z^{(k)}$ so that $(z^{(k)}, Z_\perp^{(k)})$ is orthogonal. Then

$$\begin{pmatrix} (z^{(k)})^T \\ (Z_\perp^{(k)})^T \end{pmatrix} M_k \left(z^{(k)}, Z_\perp^{(k)}\right) = \begin{pmatrix} \mu_1^{(k)} & f_k^T \\ 0 & C_k \end{pmatrix}, \tag{5.5}$$

where $f_k^T = (z^{(k)})^T M_k Z_\perp^{(k)}$ and $C_k = (Z_\perp^{(k)})^T M_k Z_\perp^{(k)}$. Since $\mu_1^{(k)}$ is a simple eigenvalue of $M_k$,

$$\mathrm{sep}(\mu_1^{(k)}, C_k) = \|(C_k - \mu_1^{(k)} I)^{-1}\|^{-1} > 0. \tag{5.6}$$

Furthermore, since $\mu_1 - \mu_1^{(k)} = \lambda_{opt} - \lambda_k \ge 0$, $\lambda_k \to \lambda_{opt}$, and

$$\mathrm{sep}(\mu_1^{(k)}, C_k) + |\mu_1 - \mu_1^{(k)}| \ge \mathrm{sep}(\mu_1, C_k) \ge \mathrm{sep}(\mu_1^{(k)}, C_k) - |\mu_1 - \mu_1^{(k)}|, \tag{5.7}$$

we must have $\mathrm{sep}(\mu_1, C_k) > 0$ and $\mathrm{sep}(\mu_1, C_k) \approx \mathrm{sep}(\mu_1^{(k)}, C_k)$ for $k$ sufficiently large.

In our notation, it is direct from Theorem 3.2 in [18] to obtain the following result.

LEMMA 5.3 ([18]). *With the previous notation, let* $\varepsilon_k = \sin\angle(y, \widetilde{\mathcal{S}}_k)$, *and assume that* $\mathrm{sep}(\mu_1, C_k) > 0$. *Then*

$$\sin\angle(y^{(k)}, y) \le \left(1 + \frac{\|M\|}{\sqrt{1-\varepsilon_k^2}\,\mathrm{sep}(\mu_1, C_k)}\right)\varepsilon_k. \tag{5.8}$$

Combining (5.3) and (5.8) with (4.29) yields the following result immediately.

THEOREM 5.4. *Suppose that* $\|s_{opt}\| = \|s_k\| = \Delta$ *and* $\mathrm{sep}(\mu_1, C_k) > 0$. *Then*

$$\sin\angle(s_k, s_{opt}) \le c_k \left(1 + \frac{\|M\|}{\sqrt{1-\varepsilon_k^2}\,\mathrm{sep}(\mu_1, C_k)}\right)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k+1}, \tag{5.9}$$

*where*

$$c_k = 2 + \frac{16(\alpha_1 + \lambda_{opt})}{(\alpha_1 - \alpha_n)^2(1-t^2)}\left(1 + \frac{k+2}{|\ln t|}\right)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2.$$

In applications, in terms of (5.7), we can approximately estimate $\mathrm{sep}(\mu_1, C_k)$ by the easily and cheaply computable $\mathrm{sep}(\mu_1^{(k)}, C_k) = \sigma_{\min}(C_k - \mu_1^{(k)} I)$, the smallest singular value of $C_k - \mu_1^{(k)} I$, by noticing that the Schur decomposition of the small sized $M_k$ can be computed efficiently so as to deliver a specific (5.5).

Finally, we establish a priori bounds for $\|(A + \lambda_k I)s_k + g\|$.

THEOREM 5.5. *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then asymptotically*

$$\|(A + \lambda_k I)s_k + g\| \leq (\lambda_{opt} - \lambda_k)\Delta + (\alpha_1 + \lambda_{opt})\|s_{opt} - s_k\|. \tag{5.10}$$

*Proof.* From (1.3), we have

$$\begin{aligned} 0 = (A + \lambda_{opt} I)s_{opt} + g &= (A + \lambda_k I + (\lambda_{opt} - \lambda_k)I)(s_k + s_{opt} - s_k) + g \\ &= (A + \lambda_k I)s_k + g + (\lambda_{opt} - \lambda_k)s_k \\ &\quad + (A + \lambda_k I)(s_{opt} - s_k) + (\lambda_{opt} - \lambda_k)(s_{opt} - s_k). \end{aligned}$$

Therefore, from $\|s_k\| = \Delta$ and $\lambda_{opt} - \lambda_k \geq 0$, noting that $\|A + \lambda_{opt} I\| = \alpha_1 + \lambda_{opt}$, we obtain

$$\begin{aligned} \|(A + \lambda_k I)s_k + g\| &= \|(\lambda_{opt} - \lambda_k)s_k + (A + \lambda_{opt} I)(s_{opt} - s_k)\| + (\lambda_{opt} - \lambda_k)\|s_{opt} - s_k\| \\ &\leq (\lambda_{opt} - \lambda_k)\Delta + \|A + \lambda_{opt} I\|\|s_{opt} - s_k\| + (\lambda_{opt} - \lambda_k)\|s_{opt} - s_k\|, \end{aligned}$$

which proves (5.10) by dropping the smaller term $(\lambda_{opt} - \lambda_k)\|s_{opt} - s_k\|$. ∎

Next we derive an a priori bound for the right-hand side of (5.10). To this end, as a by-product, by exploiting some of the previous results, it is easy to derive the following a priori bound for $\|s_k - s_{opt}\|$.

THEOREM 5.6. *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$. *Then*

$$\|s_k - s_{opt}\| \leq 4\sqrt{\kappa}\Delta \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{k+1}. \tag{5.11}$$

*Proof.* It follows from [33, Theorem 4.3] and (4.45) that

$$\|s_k - s_{opt}\| \leq 2\sqrt{\kappa}\|(I - \pi_k)s_{opt}\|,$$

where $\pi_k$ is the orthogonal projector onto $\mathcal{K}_k(A, g)$. Bound (5.11) then follows from the above relation and (6) directly. ∎

Bound (5.11) is slightly worse and clearer than (4.18b) in [33]. By substituting bound (4.46) for $\lambda_{opt} - \lambda_k$ and bound (5.11) for $\|s_k - s_{opt}\|$ into (5.10), it is straightforward to obtain the following bound for $\|(A + \lambda_k I)s_k + g\|$.

THEOREM 5.7. *Suppose* $\|s_{opt}\| = \|s_k\| = \Delta$, $k = 0, 1, \ldots, k_{\max}$, *and define* $\|r_k^{\mathrm{GLTR}}\| = \|(A + \lambda_k I)s_k + g\|$. *Then asymptotically*

$$\begin{aligned} \|r_k^{\mathrm{GLTR}}\| \leq &\left(\frac{4\eta_{k1}\Delta^2}{\beta_0} + 8(\alpha_1 + \lambda_{opt})\eta_{k2}\Delta^3\right)\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2(k+1)} \\ &+ 4\sqrt{\kappa}\Delta(\alpha_1 + \lambda_{opt})\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{k+1} \end{aligned} \tag{5.12}$$

*with the constants* $\eta_{k1}$ *and* $\eta_{k2}$ *defined by* (4.36) *and* (4.37), *respectively.*

*Remark* 5.1. The second term of the right-hand side in (5.12) dominates the bound as soon as $k$ increases, and $\|r_k^{\mathrm{GLTR}}\|$ decays at least as fast as $(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{k+1}$. Theorem 5.5, together with bounds (4.46) and (5.11), shows that $\|r_k^{\mathrm{GLTR}}\|$ converges as fast as $\|s_k - s_{opt}\|$.

*Remark* 5.2. Summarizing the results in sections 4–5, we see that the convergence rates of $\lambda_{opt} - \lambda_k$ and $q(s_k) - q(s_{opt})$ are the squares of those of $\sin\angle(s_k, s_{opt})$, $\|s_k - s_{opt}\|$, and $\|r_k^{\rm GLTR}\|$. Suppose that our bounds are realistic. Then they indicate that if $\|r_k^{\rm GLTR}\| = tol$, then $\|s_k - s_{opt}\| = \mathcal{O}(tol)$, $q(s_k) - q(s_{opt}) = \mathcal{O}(tol^2)$, and $\lambda_{opt} - \lambda_k = \mathcal{O}(tol^2)$ with the constants in big $\mathcal{O}(\cdot)$'s being modest. Therefore, if all four quantities are reduced to approximately the same level, then only about half of the iterations are needed for $q(s_k) - q(s_{opt}) = \mathcal{O}(tol^2)$ and $\lambda_{opt} - \lambda_k = \mathcal{O}(tol^2)$, compared with those needed for $\|(A + \lambda_k I)s_k + g\|$ and $\|s_k - s_{opt}\|$. Since the size of $\|r_k^{\rm GLTR}\|$ is used as the stopping criterion in the GLTR method, if one is concerned with the approximate size(s) of $\|s_k - s_{opt}\|$ and/or $q(s_k) - q(s_{opt})$, then a suitable tolerance *tol* can be prescribed for $\|r_k^{\rm GLTR}\|$.

**6. The convergence of the GLTR method for the general TRS (1.1).** In this section we extend the previous convergence results on the GLTR method with $B = I$ to the GLTR method for solving the general TRS (1.1) with $B \neq I$.

Recall the equivalence of (1.1) and (1.7). It is straightforward from (1.6) to obtain

$$q(s) = \widehat{q}(c), \ \|s\|_B = \|c\|. \tag{6.1}$$

Let $s_{opt}$ and $c_{opt}$ be the solutions to (1.1) and (1.7), respectively. Then by (1.6) and Theorem 1.1 it is easily checked that $s_{opt} = B^{-\frac{1}{2}} c_{opt}$ and the optimal Lagrangian multipliers $\lambda_{opt}$ for (1.1) and (1.7) remain the same.

It is easy to verify from definition (3.2) that

$$\mathcal{K}_k(B^{-1}A, B^{-1}g) = B^{-\frac{1}{2}}\mathcal{K}_k(\widehat{A}, \widehat{g}). \tag{6.2}$$

The preconditioned Lanczos process (cf. Algorithm 4.2) generates a $B$-orthonormal basis $\{q_i\}_{i=0}^k$ of $\mathcal{K}_k(B^{-1}A, B^{-1}g)$. Write $Q_k = (q_0, q_1, \ldots, q_k)$. Then we have

$$Q_k^T A Q_k = T_k, \ Q_k^T B Q_k = I, \ \beta_0 = \|B^{-\frac{1}{2}} g\|, \tag{6.3}$$

where $q_0 = B^{-1}g/\|B^{-1}g\|_B$ and $T_k$ is the $(k+1) \times (k+1)$ symmetric tridiagonal matrix. Let $s_k = Q_k h_k$ be the solution to the projected TRS

$$\min_{\|s\|_B \le \Delta, s \in \mathcal{K}_k(B^{-1}A, B^{-1}g)} q(s).$$

Then $h_k$ solves the reduced TRS (3.7) with $\beta_0 = \|B^{-\frac{1}{2}} g\|$.

Set $\widehat{Q}_k = B^{\frac{1}{2}} Q_k$. Then from (1.6) the first column $\widehat{q}_0$ of $\widehat{Q}_k$ is

$$\widehat{q}_0 = B^{\frac{1}{2}} q_0 = B^{-\frac{1}{2}} g/\|B^{-1}g\|_B = B^{-\frac{1}{2}} g/\|B^{-\frac{1}{2}} g\| = \widehat{g}/\|\widehat{g}\|.$$

It follows from (6.2), (6.3), and (1.6) that the columns of $\widehat{Q}_k$ form an orthonormal basis of $\mathcal{K}_k(\widehat{A}, \widehat{g})$ and

$$\widehat{Q}_k^T \widehat{A} \widehat{Q}_k = T_k, \ \beta_0 = \|B^{-\frac{1}{2}} g\| = \|\widehat{g}\|.$$

By exploiting the preconditioned Lanczos process and the symmetric Lanczos process, respectively, the above shows that solving the projected TRS

$$\min_{\|s\|_B \le \Delta, s \in \mathcal{K}_k(B^{-1}A, B^{-1}g)} q(s) \ \text{ and } \min_{\|c\| \le \Delta, c \in \mathcal{K}_k(\widehat{A}, \widehat{g})} \widehat{q}(c)$$

amounts to solving TRS (3.7). Let $c_k$ solve $\min_{\|c\| \le \Delta, c \in \mathcal{K}_k(\widehat{A}, \widehat{g})} \widehat{q}(c)$. Then

$$s_k = Q_k h_k,\ c_k = \widehat{Q}_k h_k,\ s_k = B^{-\frac{1}{2}} c_k$$

with $h_k$ the solution to the reduced TRS (3.7). Therefore, the GLTR method applied to (1.1) and (1.7) has the same optimal Lagrangian multipliers $\lambda_k$.

Suppose that TRS (1.1), i.e., TRS (1.7), is in the easy case and $\|c_{opt}\| = \Delta$ and $\|c_k\| = \Delta$ for $k \le k_{\max}$ with $k_{\max}$ the first iteration at which the symmetric Lanczos process breaks down. Then all the bounds obtained previously are directly applicable to $\lambda_{opt} - \lambda_k$, $\sin\angle(c_k, c_{opt})$, $\|c_k - c_{opt}\|$, $\widehat{q}(c_k) - \widehat{q}(c_{opt})$, and $\|(\widehat{A} + \lambda_k I)c_k + \widehat{g}\|$. In corresponding bounds we only need to replace the eigenvalues $\alpha_i$ of $A$ by the eigenvalues $\widehat{\alpha}_i$ of $\widehat{A}$, labeled as $\widehat{\alpha}_1 \ge \widehat{\alpha}_2 \ge \cdots \ge \widehat{\alpha}_n$, and the condition number $\kappa = \frac{\alpha_1 + \lambda_{opt}}{\alpha_n + \lambda_{opt}}$ by $\widehat{\kappa} = \frac{\widehat{\alpha}_1 + \lambda_{opt}}{\widehat{\alpha}_n + \lambda_{opt}}$.

By (1.6), (6.1), and the above description, we obtain

$$\|s_k - s_{opt}\|_B = \|c_k - c_{opt}\|, \tag{6.4}$$

$$q(s_k) - q(s_{opt}) = \widehat{q}(c_k) - \widehat{q}(c_{opt}), \tag{6.5}$$

$$\|(A + \lambda_k B)s_k + g\|_{B^{-1}} = \|(\widehat{A} + \lambda_k I)c_k + \widehat{g}\|, \tag{6.6}$$

where the vector $B^{-1}$-norm is defined by $\|s\|_{B^{-1}} = (s^T B^{-1} s)^{\frac{1}{2}}$. The size of $\|(A + \lambda_k B)s_k + g\|_{B^{-1}}$ is used as the stopping criterion in the GTLR method [8, 9].

For TRS (1.1), Theorem 1.1 and the symmetric Lanczos process use the $B$-inner product $(x, y)_B = x^T B y$. Correspondingly, the angle $\angle(s_k, s_{opt})_B$ of $s_k$ and $s_{opt}$ is defined via

$$\cos\angle(s_k, s_{opt})_B = \frac{(s_k, s_{opt})_B}{\|s_k\|_B \|s_{opt}\|_B} = \frac{s_k^T B s_{opt}}{\|s_k\|_B \|s_{opt}\|_B}. \tag{6.7}$$

Under the assumption $\|s_k\|_B = \|s_{opt}\|_B = \Delta$, for $\angle(s_k, s_{opt})_B$ fairly small we obtain

$$\frac{\|s_k - s_{opt}\|_B^2}{\Delta^2} = 4\sin^2\frac{\angle(s_k, s_{opt})_B}{2} \approx \sin^2\angle(s_k, s_{opt})_B.$$

From $s_k = B^{-\frac{1}{2}} c_k$, $s_{opt} = B^{-\frac{1}{2}} c_{opt}$, $\|s_k\|_B = \|c_k\|$, and $\|s_{opt}\|_B = \|c_{opt}\|$, we have

$$\cos\angle(s_k, s_{opt})_B = \cos\angle(B^{-\frac{1}{2}} c_k, B^{-\frac{1}{2}} c_{opt})_B = \frac{c_k^T c_{opt}}{\|c_k\|\|c_{opt}\|} = \cos\angle(c_k, c_{opt}),$$

so that

$$\sin\angle(s_k, s_{opt})_B = \sin\angle(c_k, c_{opt}). \tag{6.8}$$

By (6.4)–(6.6) and (6.8) and the bounds for their right-hand sides, bounds for $\|s_k - s_{opt}\|_B$, $q(s_k) - q(s_{opt})$, $\sin\angle(s_k, s_{opt})_B$, and $\|(A + \lambda_k B)s_k + g\|_{B^{-1}}$ follow directly.

**7. Numerical experiments.** In this section, we compare our a priori bounds with the true errors $\lambda_{opt} - \lambda_k$, $\sin\angle(s_k, s_{opt})$, $q(s_k) - q(s_{opt})$ and the residual norm $\|(A + \lambda_k I)s_k + g\|$, respectively. In order to give a full justification on our a priori bounds, we test our results on GLTR applied to solving TRS's in the *easy* case with $A$ having different representative eigenvalue distributions and various condition numbers $\kappa$'s. Specifically, we consider three indefinite $A$'s: the positive and negative eigenvalues of $A$ are arithmetic sequences, the extreme eigenvalues of $A$ are clustered, and the large eigenvalues of $A$ are well separated, respectively.

All the experiments were performed on an Intel Core i7, CPU 3.6GHz, 8 GB RAM using MATLAB 2017A with machine precision $\epsilon_{\mathrm{mach}} = 2.22 \times 10^{-16}$ under Microsoft Windows 10 64 bit.

In Examples 1–3, we take $n = 10000$ and a fixed trust-region radius $\Delta = 1$, and the vector $g$ is a unit length vector generated by the MATLAB built-in function $\mathsf{randn}(\mathsf{n}, 1)$. For each example, we have also tested a few different $\Delta$ and observed similar phenomena; the only distinction is that the three errors and the residual norm as well as their bounds decrease more slowly for $\Delta$ larger, as is expected from all the bounds for them since a larger $\Delta$ makes the constants in the bounds become larger. Due to such similarities and paper length, we do not report the numerical results on different $\Delta$'s. Since the uncomputable $\varepsilon_k = \sin\angle(y, \widetilde{\mathcal{S}}_k)$ tends to zero as $k$ increases, we take $\varepsilon_k = 0$ in the denominator of the bound of Theorem 5.4; we also replace $\mathrm{sep}(\mu_1, C_k)$ by its reasonable approximation $\mathrm{sep}(\mu_1^{(k)}, C_k)$. We exploit the MATLAB functions $\mathsf{eigs}$ and $\mathsf{svds}$ with the stopping tolerance $10^{-14}$ to compute $\lambda_{opt}$, $s_{opt}$, and $\|M\|$, respectively, use them as the "exact" ones, and then compute $q(s_{opt})$. To maintain the numerical orthogonality of the Lanczos basis vectors, in finite precision arithmetic, we use the symmetric Lanczos process with complete reorthogonalization.

When assessing our a priori bounds, we should note that the bounds may be large overestimates of the true errors but that there are cases where the actual errors and their bounds become close to each other when $k$ increases. Possible overestimates of our bounds are not surprising, since the bounds are established in the worst case and the constants in front of $(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{k+1}$ or $(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2(k+1)}$ are the largest possible. Our aim consists in justifying that (i) a priori bounds indeed yield sharp estimates of the asymptotic *convergence rates* even if those constants in front of the bounds are large, that is, we are concerned with the *insight* into the convergence rates, (ii) the bounds estimate the true errors and residual norms quite accurately, (iii) the practically uncomputable true errors $\lambda_{opt} - \lambda_k$, $q(s_k) - q(s_{opt})$ and their bounds are approximately the square of the computable residual norm $\|(A + \lambda_k I)s_k + g\|$ and its bound, respectively, and (iv) the true uncomputable $\sin\angle(s_k, s_{opt})$ decreases at the same rate as that of $\|(A + \lambda_k I)s_k + g\|$.

*Example* 1. This example is randomly generated, where the symmetric indefinite sparse matrix $A$ is generated by the MATLAB function

$$\mathsf{A} = \mathsf{sprandsym}(\mathsf{n}, \mathsf{density}, \mathsf{rc}), \tag{7.1}$$

where $\mathsf{rc}$ is a vector of $A$'s eigenvalues. We take $\mathsf{density} = 0.01$ and $\mathsf{rc}$ as an even distribution among $[-2, 2]$:

$$\mathsf{rc}(i) = \begin{cases} -2 + \dfrac{4}{n}(i-1), & i \le \dfrac{n}{2}, \\ 2 - \dfrac{4}{n}(n-i), & i > \dfrac{n}{2}. \end{cases}$$

*Example* 2. We take $A$ to be diagonal with translated Chebyshev nodes on the diagonal. This problem has been tested in [33]. The zero nodes of the $n$th degree Chebyshev polynomial in $[-1, 1]$ are given by

$$t_{jn} = \cos\frac{(2j-1)\pi}{2n}, \ 1 \le j \le n.$$

Given an interval $[a, b]$, the linear transformation

$$y = \left(\frac{b-a}{2}\right)\left(x + \left(\frac{a+b}{b-a}\right)\right)$$

maps $x \in [-1,1]$ to $y \in [a,b]$. The $n$th degree translated Chebyshev zero nodes on $[a,b]$ are

$$t_{jn}^{[a,b]} = \left(\frac{b-a}{2}\right)\left(t_{jn} + \left(\frac{a+b}{b-a}\right)\right),$$

which monotonically decreases for $j = 1, 2, \ldots, n/2$ and increases for $j = n/2, \ldots, n$, respectively. We take $A = \mathrm{diag}\{t_{jn}^{[a,b]}\}$, $j = 1, 2, \ldots, n$, whose eigenvalues are clustered at the two extreme points of $[a,b] = [-5,5]$.

*Example* 3. We use the Strakoš matrix $A$ [20, p. XV], which is diagonal with the eigenvalues

$$\alpha_i = \alpha_1 + \left(\frac{i-1}{n-1}\right)(\alpha_n - \alpha_1)\rho^{n-i},$$

$i = 1, 2, \ldots, n$. The parameter $\rho$ controls the eigenvalue distribution. The large eigenvalues of $A$ are better separated for $\rho < 1$ smaller. We take $\alpha_1 = 8$, $\alpha_n = -2$, and $\rho = 0.99$ in the experiment.

We list the results of Example 1 in Table 1 and draw our bounds and the true $\lambda_{opt} - \lambda_k$, $\sin\angle(s_k, s_{opt})$, $q(s_k) - q(s_{opt})$, and $\|(A + \lambda_k I)s_k + g\|$ in Figure 1, where the recorded $k$ denotes the iteration, after which the corresponding errors or residual norms do not decrease further, i.e., they stagnate or stabilize if GLTR is performed for more iterations. This meaning applies to later experiments as well.

We have also made experiments on Examples 2–3 and found $\kappa = 34.9455$ and 11.1518, respectively. We have observed phenomena similar to those for Example 1. Due to the similarity and the paper length, we omit details. Rather, we will make several comments on the results of Examples 1–3 and get insight into common features and subtle distinctions.

For the numerical results on Examples 1–3, we have the following observations: (i) the corresponding bounds predict the convergence rates of $\lambda_{opt} - \lambda_k$, $\sin\angle(s_k, s_{opt})$, $q(s_k) - q(s_{opt})$, and $\|(A+\lambda_k I)s_k + g\|$ accurately, (ii) the bounds are very close to their values in most of the cases, especially for $\lambda_{opt} - \lambda_k$, $q(s_k) - q(s_{opt})$ and $\|(A+\lambda_k I)s_k + g\|$, and (iii) $\lambda_k > 0$ happens when $k = 1$ for all the test problems, as is clearly seen from Figure 1(a) by observing that $\lambda_{opt} - \lambda_k$ *strictly* decreases monotonically from $k = 1$ upward. In contrast, relatively speaking, the bound for $\sin\angle(s_k, s_{opt})$ is not so sharp,

TABLE 1
*Example* 1*, experiments with the easy case.*

Parameters in Example 1.

| $\alpha_1$ | $\alpha_n$ | $\kappa$ | $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ | $\lambda_{opt}$ | $q(s_{opt})$ |
|---|---|---|---|---|---|
| 2.0000 | −2.0000 | 18.1481 | 0.6198 | 2.2333 | −1.4770 |

$\lambda_{opt} - \lambda_k$, $q(s_k) - q(s_{opt})$, $\sin\angle(s_k, s_{opt})$, $\|(A+\lambda_k I)s_k + g\|$ and their bounds (4.46), (4.43), (5.9), (5.12).

| $k$ | $\lambda_{opt} - \lambda_k$ | bound | $k$ | $q(s_k) - q(s_{opt})$ | bound |
|---|---|---|---|---|---|
| 34 | $1.07e-13$ | $2.67e-13$ | 34 | $3.33e-15$ | $2.52e-13$ |
| $k$ | $\sin\angle(s_k, s_{opt})$ | bound | $k$ | $\|(A+\lambda_k I)s_k + g\|$ | bound |
| 67 | $1.82e-14$ | $5.46e-11$ | 66 | $1.89e-14$ | $1.40e-12$ |

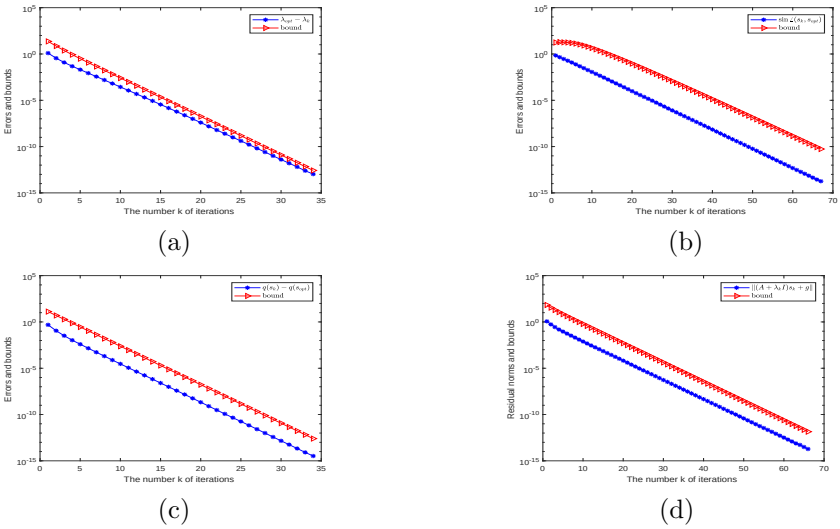


FIG. 1. *Example* 1, *experiments with the easy case.* (a) $\lambda_{opt} - \lambda_k$ *and its bound* (4.46)*;* (b) $\sin\angle(s_k, s_{opt})$ *and its bound* (5.9)*;* (c) $q(s_k) - q(s_{opt})$ *and its bound* (4.43)*;* (d) $\|(A + \lambda_k I)s_k + g\|$ *and its bound* (5.12)*.*

yet it predicts the convergence rate of $\sin\angle(s_k, s_{opt})$ accurately. The reason should be due to possible small $\mathrm{sep}(\mu_1^{(k)}, C_k)$. This may be a disadvantage of transforming (3.1) and (3.7) into the mathematically equivalent matrix eigenvalue problem of $M_k$ defined by (4.3). Two equivalent mathematical problems have different condition numbers for $\|s_k - s_{opt}\|$ and $\sin\angle(s_k, s_{opt})$, and $s_k$ may be worse conditioned from the viewpoint of matrix eigenvalue problems than it is from the viewpoint of GLTR itself.

In the meantime, our numerical results have also confirmed that (i) $\lambda_{opt} - \lambda_k$ and $q(s_k) - q(s_{opt})$ as well as their bounds are approximately the squares of $\sin\angle(s_k, s_{opt})$ and $\|(A + \lambda_k I)s_k + g\|$, (ii) the former two and their bounds use approximately half of the iterations needed for the latter two and their bounds to achieve approximately the same tolerance, and (iii) the condition number $\kappa$ affects the convergence of the GLTR method: the bigger $\kappa$ is, the more iterations the method needs to reduce each of $\lambda_{opt} - \lambda_k$, $\sin\angle(s_k, s_{opt})$, $q(s_k) - q(s_{opt})$, and $\|(A + \lambda_k I)s_k + g\|$ to approximately the same level. This confirms Remark 5.2. Therefore, we can use the size of the computable $\|(A + \lambda_k I)s_k + g\|$ to predict the other three uncomputable errors. Also, recall from Remark 5.1 that $\|(A+\lambda_k I)s_k+g\|$ is as small as $\|s_k - s_{opt}\|$. Consequently, all these imply that if one is concerned with the size of the uncomputable error $q(s_k) - q(s_{opt})$ of the optimal objective value or that of the uncomputable $\|s_k - s_{opt}\|$, then we can prescribe the stopping tolerance for $\|(A + \lambda_k I)s_k + g\|$.

It is important to stress that we claim all the bounds for the three errors and the residual norm to be sharp under the implicit *assumption* that bounds (4.16) and (4.20) are sharp. We have then derived sharp estimates for $\epsilon_1^{(k)}$ and $\epsilon_2^{(k)}$, respectively. Precisely, such implicit assumption means that $g$ has roughly comparable components in the directions of all the eigenvectors $v_i$ of $A$. Expand $g$ in the basis $\{v_i\}_{i=1}^n$ as $g = \sum_{i=1}^n \beta_i v_i$. Then

$$\beta_i = v_i^T g = \|g\| \cos\angle(g, v_i), \ \sum_{i=1}^{n} \beta_i^2 = \|g\|^2,$$

meaning that $\sum_{i=1}^{n} \cos^2 \angle(g, v_i) = 1$. Therefore, if $g$ is general, that is, it has roughly equal sized components in all the $v_i$, i.e., all the $|\beta_i|$ are roughly equal, then we roughly have $|\cos \angle(g, v_i)| \approx \frac{1}{\sqrt{n}}$. Suppose that $\alpha_n$ is simple. Then $v_n$ spans $\mathcal{N}(A - \alpha_n I)$. If $|\cos \angle(g, v_n)|$ is *considerably* smaller than $\frac{1}{\sqrt{n}}$, bounds (4.16) and (4.20) are *considerable* overestimates, and bound (4.18) for $\epsilon_1^{(k)}$ is too, causing our bounds for $\lambda_{opt} - \lambda_k$, $q(s_k) - q(s_{opt})$, $\|s_k - s_{opt}\|$, and $\|r_k^{\mathrm{GLTR}}\|$ to be substantial overestimates. In this case, one may have a large $\kappa = \frac{\alpha_1 + \lambda_{opt}}{\alpha_n + \lambda_{opt}}$, as it is straightforwardly obtained from Theorem 4.7 of [33] that

$$\kappa \leq \frac{1}{\mathcal{O}(|\cos \angle(g, v_n)|)} \tag{7.2}$$

with the constant in $\mathcal{O}(\cdot)$ modest for modest $\Delta$ and $\beta_0$. A large $\kappa$ must mean a small $|\cos \angle(g, v_n)|$, but the converse may be untrue. In finite precision arithmetic, if $|\cos \angle(g, v_n)| = \mathcal{O}(\epsilon_{\mathrm{mach}})$ and $\kappa = \mathcal{O}(\frac{1}{\epsilon_{\mathrm{mach}}})$, then TRS is in the hard case. At this time, as we have elaborated in section 3, GLTR itself cannot solve the TRS in the hard case. However, if $|\cos \angle(g, v_n)|$ is only fairly small and $\kappa$ is fairly large, e.g., $|\cos \angle(g, v_n)| = \mathcal{O}(10^{-5})$ and $\kappa = \mathcal{O}(10^5)$, then TRS is numerically in the easy case.

We next investigate the behavior of GLTR and our bounds for relatively small $|\cos \angle(g, v_n)|$ and large $\kappa$'s. Given a unit length vector $g$ generated randomly in a normal distribution, we orthogonalize it against $v_n$ to obtain

$$g_\perp = (I - v_n v_n^T) g.$$

Adding a perturbation vector $\varepsilon v_n$ to it, we construct the new vector $g(\varepsilon)$ in TRS:

$$g(\varepsilon) = g_\perp + \varepsilon \|g_\perp\| v_n, \ g(\varepsilon) := g(\varepsilon)/\|g(\varepsilon)\|, \tag{7.3}$$

so that $\cos \angle(g(\varepsilon), v_n) = \frac{\varepsilon}{(1+\varepsilon^2)^{1/2}} \approx \varepsilon$. We then adjust $\Delta$ to make $\kappa = \mathcal{O}(\frac{1}{\varepsilon})$.

For Example 3 with $n = 10000$, taking $\varepsilon = 10^{-4} \ll \frac{1}{\sqrt{n}} = 10^{-2}$ and $\Delta = 1$, we found that $\kappa = 9.9 \times 10^4$, and $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 0.9937$, $\lambda_{opt} = 2.0000$, and $q(s_{opt}) = -1601.51$. Table 2 lists the results, and Figure 2 draws the convergence processes of four quantities and their bounds. Clearly, the GLTR converges very slowly in the first 35 iterations, and the bounds are meaningless. Particularly, unlike the case that $g$ is a general vector, bound (5.9) for $\sin \angle(s_k, s_{opt})$ does not decrease in the first several iterations because of slowly increasing big multiples in the bound. Moreover, Figure 2 illustrates that the decreasing rates of true errors and residual norms are not uniform any longer and each of them decays at different rates in different stages as iterations proceed. The residual norms may even not decrease monotonically and finally stabilize around $10^{-8}$, i.e., $\kappa\mathcal{O}(\sqrt{n}\|A\|\epsilon_{\mathrm{mach}})$.

We have found that GLTR behaves similarly for several other $\varepsilon \in [10^{-6}, 10^{-4}]$ and $\kappa \in [10^4, 10^6]$. For Examples 1–2, we have observed similar phenomena. For instance, in Example 1, by taking $\Delta = 40$, $\varepsilon = 10^{-4}$, and $\kappa = 8.1 \times 10^5$, the residual norms decay very slowly and stabilize around $6.0 \times 10^{-7} = \kappa\mathcal{O}(\sqrt{n}\|A\|\epsilon_{\mathrm{mach}})$ after 600 iterations, much more than those for a general $g$. It is instructive to observe that GLTR does not converges as fast as it does for a general $g$, and ultimately attainable residual norms and $\sin \angle(s_k, s_{opt})$ are dictated by the size of $\kappa$.

Obviously, if $\kappa$ is relatively large but TRS is numerically in the easy case, the convergence of GLTR needs a separate analysis.

TABLE 2
*Example 3, experiments with a harder case.*

| $\lambda_{opt}-\lambda_k$, $q(s_k)-q(s_{opt})$, $\sin\angle(s_k,s_{opt})$, $\|(A+\lambda_k I)s_k+g\|$ and their bounds (4.46), (4.43), (5.9), (5.12). | | | | | |
|---|---|---|---|---|---|
| $k$ | $\lambda_{opt}-\lambda_k$ | bound | $k$ | $q(s_k)-q(s_{opt})$ | bound |
| 60 | $2.03e-12$ | 75.6340 | 65 | $4.89e-15$ | 34.9650 |
| $k$ | $\sin\angle(s_k,s_{opt})$ | bound | $k$ | $\|(A+\lambda_k I)s_k+g\|$ | bound |
| 70 | $8.69e-09$ | $3.95e+10$ | 70 | $6.56e-09$ | $8.11e+03$ |

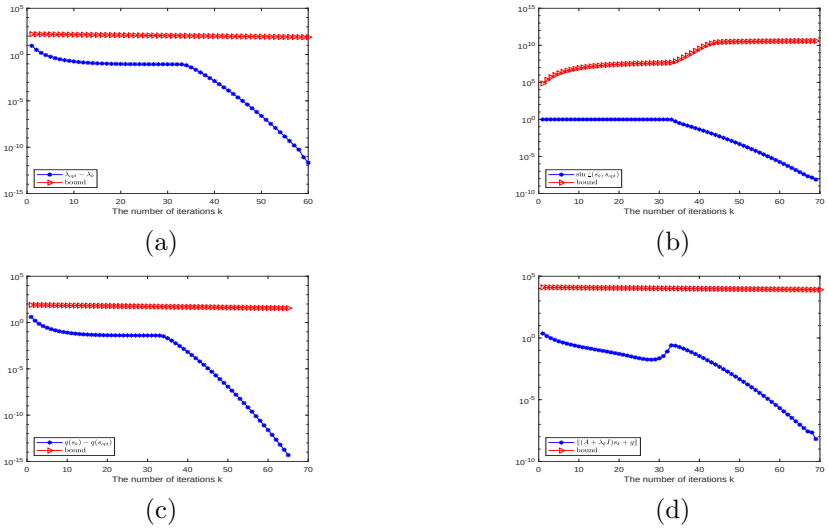


FIG. 2. *Example 3, experiments with the harder case* $g(10^{-4})$. (a) $\lambda_{opt}-\lambda_k$ *and its bound;* (b) $\sin\angle(s_k,s_{opt})$ *and its bound;* (c) $q(s_k)-q(s_{opt})$ *and its bound;* (d) $\|(A+\lambda_k I)s_k+g\|$ *and its bound.*

**8. Conclusion.** The GLTR method has received much attention for years. For TRS (1.1) in the easy case, there has been no quantitative analysis and result on $\lambda_{opt}-\lambda_k$ and $\|(A+\lambda_k I)s_k+g\|_{B^{-1}}$. Starting with the mathematical equivalence of the solution of TRS (1.7) with $B=I$ and the eigenvalue problem of the matrix $M$ defined by (2.3), we have established a priori bounds for $\lambda_{opt}-\lambda_k$, $\sin\angle(s_k,s_{opt})$, $q(s_k)-q(s_{opt})$, and the residual norm $\|(A+\lambda_k I)s_k+g\|$. We have extended all the convergence results to the GLTR method for solving the general TRS (1.1) with $B\neq I$.

The bound for $\|(A+\lambda_k I)s_k+g\|$ is of particular importance in theory and practice, and it can predict the uncomputable errors $\lambda_{opt}-\lambda_k$, $\|s_k-s_{opt}\|$, and $q(s_k)-q(s_{opt})$. If one is interested in approximately reducing $\|s_k-s_{opt}\|$ and/or $q(s_k)-q(s_{opt})$ to some level, then a stopping tolerance can be prescribed for $\|(A+\lambda_k I)s_k+g\|$.

Numerical results have confirmed that for a general $g$, our bounds are realistic and predict the convergence rates of the three errors and the residual norm in the GLTR method accurately. On the other hand, for the case that $g$ is fairly close to $\mathcal{N}(A-\alpha_n I)$ and $\kappa$ is relatively large, a separate analysis is appealing.

**Acknowledgment.** We thank two referees very much for their valuable suggestions and comments, which helped us to improve the presentation considerably.

## REFERENCES


[1] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda, *Solving the trust-region subproblem by a generalized eigenvalue problem*, SIAM J. Optim., 27 (2017), pp. 269–291.

[2] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *Approximate solution of the trust region problem by minimization over two-dimensional subspaces*, Math. Progam., 40 (1988), pp. 247–263.

[3] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MOS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[4] R. Ei Attar, *Special Functions and Orthogonal Polynomials*, Lulu Press, 2006.

[5] J. B. Erway and P. E. Gill, *A subspace minimization method for the trust-region step*, SIAM J. Optim., 20 (2010), pp. 1439–1461.

[6] J. B. Erway, P. E. Gill, and J. D. Griffin, *Iterative methods for finding a trust-region step*, SIAM J. Optim., 20 (2009), pp. 1110–1131.

[7] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw., 19 (2004), pp. 41–67.

[8] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.

[9] N. I. M. Gould, D. Orban, and P. L. Toint, *GALAHAD, a library of thread-safe Fortran* 90 *packages for large-scale nonlinear optimization*, ACM Trans. Math. Software, 29 (2004), pp. 353–372.

[10] N. I. M. Gould, D. P. Robinson, and H. S. Thorne, *On solving trust-region and other regularised subproblems in optimization*, Math. Program. Comput., 2 (2010), pp. 21–57.

[11] W. Gander, C. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114 (1989), pp. 815–839.

[12] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers in Appl. Math. 17, SIAM, Philadephia, 1997.

[13] T. M. Huang, Z. Jia, and W. W. Lin, *On the convergence of Ritz pairs and refined Ritz vectors for quadratic eigenvalue problems*, BIT, 53 (2013), pp. 941–958.

[14] W. W. Hager, *Minimizing a quadratic over a sphere*, SIAM J. Optim., 12 (2001), pp. 188–208.

[15] W. W. Hager and Y. Krylyuk, *Graph partitioning and continuous quadratic programming*, SIAM J. Algebra Discrete Methods, 12 (1999), pp. 500–523.

[16] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[17] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 843–862.

[18] Z. Jia and G. W. Stewart, *On the Convergence of Ritz Values, Ritz Vectors and Refined Ritz Vectors*, Tech. report TR-3986, Department of Computer Science, University of Maryland, College Park, 1999.

[19] L. Lukšan, C. Matonoha, and J. Vlček, *On Lagrange multipliers of trust-region subproblems*, BIT, 48 (2008), pp. 763–768.

[20] G. Meurant, *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, Software Environ. Tools 19, SIAM, Philadelphia, 2006.

[21] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.

[22] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

[23] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Appl. Math. 20, SIAM, Philadephia, 1998.

[24] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Math. Program., 77 (1997), pp. 273–299.

[25] M. Rojas, S. A. Santos, and D. C. Sorensen, *A new matrix-free algorithm for the large-scale trust-region subproblem*, SIAM J. Optim., 11 (2001), pp. 611–646.

[26] M. Rojas, S. A. Santos, and D. C. Sorensen, *Algorithm* 873*: LSTRS: MATLAB software for large-scale trust-region subproblems and regularization*, ACM Trans. Math. Software, 34 (2008), pp. 1–28.

[27] Y. Saad, *Numerical Methods for Large Eigenvalue Problems: Revised Edition*, Classics in Appl. Math. 66, SIAM, Philadephia, 2011.

[28] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.

[29] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[30] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[31] P. L. Toint, *Towards an Efficient Sparsity Exploiting Newton Method for Minimization*, in Sparse Matrices and Their Uses, Academic Press, New York, 1981, pp. 57–88.

[32] Y. Yuan, *On the truncated conjugate gradient method*, Math. Program., 87 (2000), pp. 561–573.
[33] L. H. Zhang, C. G. Shen, and R. C. Li, *On the generalized Lanczos trust-region method*, SIAM J. Optim., 27 (2017), pp. 2110–2142.
[34] L. H. Zhang, W. H. Yang, C. Shen, and J. Feng, *Error bounds of the Lanczos approach for the trust-region subproblem*, Front. Math. China, 13 (2018), pp. 459–481.