On the Generalized Lanczos Trust-Region Method

Lei-Hong Zhang* Chungen Shen† Ren-Cang Li‡

March 12, 2017

Abstract

The so-called *Trust-Region Subproblem* gets its name in the trust-region method in optimization and also plays a vital role in various other applications. Several numerical algorithms have been proposed in the literature for solving small-to-medium size dense problems as well as for large scale sparse problems. The Generalized Lanczos Trust-Region (GLTR) method proposed by [Gould, Lucidi, Roma and Toint, *SIAM J. Optim.*, 9:504–525 (1999)] is a natural extension of the classical Lanczos method for the linear system to the trust-region subproblem. In this paper, we first analyze the convergence of GLTR to reveal its convergence behavior in theory and then propose new stopping criteria that can be integrated into GLTR for better numerical performance. Specifically, we develop *a priori* upper bounds for the convergence to both the optimal objective value as well as the optimal solution, and argue that these bounds can be efficiently estimated numerically and serve as stopping criteria for iterative methods such as GLTR. Two sets of numerical tests are presented. In the first set, we demonstrate the sharpness of the upper bounds, and for the second set, we integrate the upper bound estimate into the Fortran routine `GLTR` in the library `GALAHAD` as new stopping criteria, and test the trust-region solver `TRU` on the problem collection `CUTEr`. The numerical results show that, with the new stopping criteria in `GLTR`, the overall performance of `TRU` can be improved considerably.

Key words. Trust-region subproblem, Lanczos method, conjugate gradient method, Trust-region method, convergence, stopping criterion

AMS subject classifications. 90C20, 90C06, 65F10, 65F15, 65F35

*School of Mathematics and Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China. The work of this author was supported in part by the National Natural Science Foundations of China NSFC-11371102, NSFC-11671246, and the Basic Academic Discipline Program, the 11th five year plan of 211 Project for Shanghai University of Finance and Economics. Email: `zhang.leihong@mail.shufe.edu.cn`.

†College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China. Email: `shenchungen@gmail.com`. The work of the first author is supported in part by the National Natural Science Foundations of China NSFC-11101281 and NSFC-11271259.

‡Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019-0408, USA. Email: `rcli@uta.edu`. This work was supported in part by NSF grants DMS-1317330 and CCF-1527104, and by NSFC grant 11428104.

1 Introduction

Minimization of a quadratic function over a Euclidean ball

$$\min_{\|\boldsymbol{s}\|_2 \le \Delta} f(\boldsymbol{s}) \quad \text{with} \quad f(\boldsymbol{s}) := \frac{1}{2}\boldsymbol{s}^\top H\boldsymbol{s} + \boldsymbol{s}^\top \boldsymbol{g} \tag{1.1}$$

is widely known as the *trust-region subproblem* (TRS) [20, 21], where $H = H^\top \in \mathbb{R}^{n\times n}$, $\boldsymbol{g} \in \mathbb{R}^n$ and $\Delta > 0$ is the trust-region radius. It also shows up in other important applications such as the Tikhonov regularization [24, 25, 26, 32, 33] for ill-posed problems, graph partitioning problems [13] as well as in the Levenberg-Marquardt approach in optimization [21].

Because of its vital role in numerous applications, several algorithms have been proposed for (1.1). Basically, these algorithms can be classified into two categories: algorithms based on matrix factorizations for small-to-medium size dense problems (see, e.g., [20, 21]) and factorization-free algorithms for large-scale sparse problems (see, e.g., [8, 12, 21, 23, 24, 25, 26, 30, 31, 32, 34]). The method proposed in [11] is based on matrix factorization but can often be just as successful as factorization-free methods for large problems when the sparsity structure is favorable.

The Moré-Sorensen method [20] available as subroutine `GQTPAR` in MINPACK-2 is probably the most well-known one for small size dense problems, and it and its modifications are frequently embedded into programs as building blocks for solving relevant subproblems within large-scale computational problems. This is also the case for the method proposed in [9] (see also [4, Chapter 5]). In particular, the authors of [9] presented a generalized Lanczos trust-region method (GLTR) [9, Algorithm 5.1] which is an improved Steihaug–Toint [31, 34] truncated Conjugate Gradient (tCG) iteration for the weighted-norm trust-region subproblem:

$$\min_{\|\boldsymbol{s}\|_M \le \Delta} f(\boldsymbol{s}), \tag{1.2}$$

where the weighting matrix $M \in \mathbb{R}^{n\times n}$ is a given symmetric positive definite matrix, and $\|\boldsymbol{s}\|_M := \sqrt{\boldsymbol{s}^\top M\boldsymbol{s}}$ is the M-vector norm of $\boldsymbol{s}$.

GLTR generally consists of two main phases, namely *the first pass* and *the second pass*. In the first pass, it starts with the (preconditioned) CG iteration [9, Algorithm 4.1] for minimizing $f(\boldsymbol{s})$. During the (preconditioned) CG iterations, the objective function decreases while the M-norm of the iterative solution increases. Thus, the iteration in the first pass stops either when the solution $\boldsymbol{s}_{\text{opt}}$ to (1.2) is achieved within the trust region $\|\boldsymbol{s}\|_M \le \Delta$, or one CG step exceeds Δ in M-norm or directions with negative curvature (a vector $\boldsymbol{p}$ is a direction of negative curvature if $\boldsymbol{p}^\top H\boldsymbol{p} < 0$) are detected. The former means that the problem (1.2) is equivalent to a linear system $H\boldsymbol{s} = -\boldsymbol{g}$, while the latter implies either H is indefinite and $\|\boldsymbol{s}_{\text{opt}}\|_M = \Delta$, or H is positive semi-definite but there is a $\boldsymbol{p} \in \mathbb{R}^n$ such that $H\boldsymbol{p} = 0$ and $\boldsymbol{p}^\top\boldsymbol{g} < 0$; the second pass will be triggered thereafter to obtain (together with a third pass) an approximation $\boldsymbol{s}_k$ on the boundary. By making use of the close relationship between CG and the Lanczos process (see [9, Section 4]), in the second pass, GLTR needs to solve smaller size trust-region subproblems successively, which are resulted from projecting the original TRS (1.2) onto the Krylov subspace generated by the Lanczos process, or equivalently, by CG (see [9] in detail). Extensive numerical

testing suggests that by integrating the first pass and the second pass carefully [9, Section 5.1], GLTR is able to achieve efficiently a boundary solution on the one hand, and also maintains the efficiencies of CG so long as the iterates lie in the interior, on the other hand.

GLTR can be understood as a generalization and indeed an efficient implementation of a kind of Lanczos method for TRS as detailed in [9, Section 5] under the name of *truncated Lanczos approach* (abbreviated as TLTRS in this paper). In particular, TLTRS mimics the classical Rayleigh-Ritz procedure (see [22, Section 11.3] and [5, Definition 7.1]) for the eigenvalue problem and proceeds iteratively the following three steps: for $k = 0, 1, \ldots$ (more detail will be given in section 3):

1. generate the kth Krylov subspace by the preconditioned Lanczos process [22, Algorithm 4.2], or equivalently the preconditioned CG [9, Algorithm 4.1];

2. project the original TRS (1.2) onto the kth Krylov subspace to give a smaller size TRS;

3. solve the resulting smaller size TRS to get an approximate solution to TRS (1.2).

TLTRS can be viewed as a natural extension of the classical Lanczos method (see e.g., [5, 28]) for the linear system and eigenvalue problem to TRS. There has been a wealth of developments, in both theory and implementation, on the Lanczos-based methods, e.g., in [5, 22, 28] for a complete development up to 1998 and more recently in [15, 16, 17]. However, to the authors' best knowledge, convergence analysis for the Lanczos type method for TRS has not yet been fully developed.

Our goals in this paper are two-fold. First, on the theoretic aspect, we analyze the convergence of TLTRS. In contrast to *a posteriori* error bounds in [36], we will develop *a priori* upper bounds for both the convergence to the optimal objective value as well as the optimal solution during TLTRS iterations. Second, on the numerical aspect, we will offer practical and effective estimates of the upper bounds. These estimates can be computed at roughly $O(k^2)$ extra flops, which turns out to be practical as $k \ll n$ in general, and therefore can be used as stopping criteria for the second pass in GLTR, an efficient implementation of TLTRS.

We conduct two sets of numerical tests to support both our theoretical bounds and their practical estimates used as effective stopping criteria. In the first set of testing, we present several numerical tests to show the sharpness of these upper bounds, and in the second set, we integrated our upper bound estimates into the Fortran package `GLTR` in the library `GALAHAD`[1] (version 2.6), and tested the trust-region method implemented in the solver `TRU` on unconstrained minimization problems with $n \geq 100$ from `CUTEr` collection (86 test problems in all). The numerical results show that, with the new stopping criteria integrated into `GLTR`, the overall performance of the trust-region solver `TRU` improves considerably.

The rest of this paper is organized as follows. In section 2, we first present some preliminary results on TRS, where the so-called the *nondegenerate* case (or *easy* case) and the *degenerate* case (or the *hard* case) are explicitly stated. In section 3, we then briefly

[1] `GALAHAD` is a thread-safe library of Fortran 2003 packages for solving nonlinear optimization problems and its version 2.6 is available at http://www.galahad.rl.ac.uk/.

describe the framework of TLTRS/GLTR as well as some basic properties. Section 4 contains the main convergence results of this paper: in subsections 4.1 and 4.2, we discuss the convergence of TLTRS for the case $\lambda_{\rm opt} = 0$ and $\lambda_{\rm opt} \neq 0$, respectively, where $\lambda_{\rm opt}$ denotes the Lagrangian multiplier of (1.1) associated with the solution $\boldsymbol{s}_{\rm opt}$; subsection 4.3 shows how to extend the main convergence results to the weighted-norm TRS (1.2). Our numerical verification of the sharpness of the established upper bounds is carried out in section 5. In section 6, we suggest a new stopping criteria for `GLTR` in `GALAHAD` and its numerical performance in comparison with the original `GLTR` is presented in section 7. Final conclusions are drawn in section 8.

Notation. Throughout this paper, all vectors are column vectors and are typeset in bold lower case letters. For $\boldsymbol{x} \in \mathbb{R}^n$ (the set of all real n-vectors), x_i stands for its ith entry. For $A \in \mathbb{R}^{m\times n}$ (the set of all $m \times n$ real matrices), $A^\dagger$ stands for the Moore-Penrose inverse of A, and $A^\top$ and $\mathcal{R}(A)$ denote its transpose and range, respectively. The $n \times n$ identity matrix is I_n or simply I if its size is clear from the context, and $\boldsymbol{e}_j$ is the jth column of an identity matrix whose size is determined by the context. To simplify our presentation, we shall also adopt MATLAB-like convention to access the entries of vectors and matrices. For example, $A_{(i,j)}$ is (i, j)th entry of A. With $i : j$ for the set of integers from i to j inclusive, $A_{(k:\ell,i:j)}$ is the sub-matrix of A that consists of intersections of row k to row ℓ and column i to column j.

2 Optimality Conditions

The following well-known optimality conditions are due to Gay [7] and Moré and Sorensen [20] (see also [29] and [21, Theorem 4.1]). It has been serving as the fundamental guideline for most existing methods for TRS.

Theorem 2.1 ([29]). *The vector $\boldsymbol{s}_{\rm opt}$ is a global optimal solution of the trust-region problem* (1.1) *if and only if $\boldsymbol{s}_{\rm opt}$ is feasible, i.e., $\|\boldsymbol{s}_{\rm opt}\|_2 \le \Delta$, and there is a scalar $\lambda_{\rm opt} \ge 0$ such that the following conditions are satisfied:*

$$\begin{cases} (H + \lambda_{\rm opt} I_n)\boldsymbol{s}_{\rm opt} = -\boldsymbol{g}, \quad \lambda_{\rm opt}(\Delta - \|\boldsymbol{s}_{\rm opt}\|_2) = 0, \quad \textit{and} \\ H + \lambda_{\rm opt} I_n \textit{ is positive semidefinite.} \end{cases}$$

Let the eigen-decomposition of H be

$$H = U\Theta U^\top \quad \text{with} \quad \Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n),$$

where the eigenvector matrix $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ is orthogonal, and

$$\theta_1 = \theta_2 = \cdots = \theta_p < \theta_{p+1} \le \cdots \le \theta_n \tag{2.1}$$

are the eigenvalues. In (2.1), we assume θ_1 has multiplicity p. Let $\mathcal{E}_1$ be the invariant subspace associated with the smallest eigenvalue θ_1. Then $U_1 = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{n\times p}$ is an orthonormal basis matrix for $\mathcal{E}_1$. Write $U = [U_1, U_2]$, where $U_2 = [\boldsymbol{u}_{p+1}, \ldots, \boldsymbol{u}_n]$, and set $\mathcal{E}_2 = \mathcal{R}(U_2) = \mathcal{E}_1^\perp$, the orthogonal complement of $\mathcal{E}_1$.

For TRS (1.1), there are two situations (see, e.g., [12, 20, 21]) to consider:

1. the *degenerate* case [12, Lemma 2.2] (or the *hard case* [21]) as characterized by[2]

$$\boldsymbol{g} \perp \mathcal{E}_1 \quad \text{and} \quad \|(H - \theta_1 I_n)^\dagger \boldsymbol{g}\|_2 \leq \Delta \tag{2.2}$$

and the corresponding Lagrangian multiplier is $\lambda_{\text{opt}} = -\theta_1$. In this case, there are multiple global solutions which can be expressed by [12, Lemma 2.2]

$$\boldsymbol{s}_{\text{opt}} = -(H - \theta_1 I_n)^\dagger \boldsymbol{g} + \tau \boldsymbol{u} \tag{2.3}$$

for any $\boldsymbol{u} \in \mathcal{E}_1$ with $\|\boldsymbol{u}\|_2 = 1$, and

$$\tau^2 = \Delta^2 - \|(H - \theta_1 I_n)^\dagger \boldsymbol{g}\|_2^2 \geq 0;$$

2. the *nondegenerate* case [12, Lemma 2.2] (or the *easy case* [21]) as characterized by the opposite of (2.2). In this case, the corresponding Lagrangian multiplier $\lambda_{\text{opt}} > \max\{-\theta_1, 0\}$, and the global solution $\boldsymbol{s}_{\text{opt}}$ is unique and given by [12, Lemma 2.2]

$$\boldsymbol{s}_{\text{opt}} = -(H + \lambda_{\text{opt}} I_n)^{-1} \boldsymbol{g}.$$

By investigating these two cases, it can be seen that if H is positive definite, the global solution $\boldsymbol{s}_{\text{opt}}$ can only be either $\boldsymbol{s}_{\text{opt}} = -H^{-1}\boldsymbol{g}$ (i.e., $\lambda_{\text{opt}} = 0$) or $\boldsymbol{s}_{\text{opt}} = -(H + \lambda_{\text{opt}} I_n)^{-1}\boldsymbol{g}$ (i.e., $\lambda_{\text{opt}} > -\theta_1$) on the boundary. Therefore, the degenerate case can only occur when $\theta_1 \leq 0$.

3 The truncated Lanczos approach for TRS (TLTRS)

We first outline TLTRS method [9, section 5]. Given a symmetric positive definite $M \in \mathbb{R}^{n\times n}$, TLTRS starts by using the generalized Lanczos process to produce an M-orthonormal basis $Q_k = [\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_k] \in \mathbb{R}^{n\times(k+1)}$ of the $(k+1)$st Krylov subspace[3]

$$\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g}) := \mathcal{R}(M^{-1}\boldsymbol{g}, (M^{-1}H)M^{-1}\boldsymbol{g}, \ldots, (M^{-1}H)^k M^{-1}\boldsymbol{g}), \quad k = 0, 1, \ldots$$

of $M^{-1}H$ on $M^{-1}\boldsymbol{g}$ to partially reduce H to the tridiagonal matrix

$$T_k := Q_k^\top H Q_k = \begin{bmatrix} \delta_0 & \gamma_1 & & & \\ \gamma_1 & \delta_1 & \gamma_2 & & \\ & \cdot & \cdot & \cdot & \\ & & \gamma_{k-1} & \delta_{k-1} & \gamma_k \\ & & & \gamma_k & \delta_k \end{bmatrix}, \tag{3.1}$$

where $Q_k^\top M Q_k = I_{k+1}$ [9, Algorithm 4.2], assuming

$$\dim \mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g}) = k + 1.$$

[2]We adopt the definitions of degenerate and nondegenerate cases used in [12, Lemma 2.2] in this paper.

[3]We adopt a notation convention that is consistent with the one used in [9]. That is to use $\mathcal{K}_k$ for the Krylov subspace of order $(k+1)$, and accordingly T_k and Q_k for the generated $(k+1)\times(k+1)$ symmetric tridiagonal matrix and $n \times (k+1)$ orthonormal basis matrix, different from $\mathcal{K}_{k+1}$, T_{k+1} and Q_{k+1} that are customarily used in the numerical linear algebra community [5, p.305].

Compactly, the process can be expressed by the relation

$$HQ_k - MQ_kT_k = \gamma_{k+1}M\boldsymbol{q}_{k+1}\boldsymbol{e}_{k+1}^\top \tag{3.2}$$

with $\gamma_0 = \|M^{-1}\boldsymbol{g}\|_2$, $Q_k\boldsymbol{e}_1 = \boldsymbol{q}_0 := \gamma_0^{-1}(M^{-1}\boldsymbol{g})$. This leads to the following reduced trust-region subproblem

$$\min_{\|\boldsymbol{h}\|_2\le\Delta} \phi(\boldsymbol{h}) \quad \text{with} \quad \phi(\boldsymbol{h}) := \frac{1}{2}\boldsymbol{h}^\top T_k\boldsymbol{h} + \gamma_0\boldsymbol{h}^\top\boldsymbol{e}_1. \tag{3.3}$$

Let $\boldsymbol{h}_k$ be the minimizer of (3.3). It can be verified that the vector

$$\boldsymbol{s}_k = Q_k\boldsymbol{h}_k \in \mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g})$$

is the minimizer of

$$\min_{\substack{\boldsymbol{s}\in\mathcal{K}_k(M^{-1}H,M^{-1}\boldsymbol{g})\\ \|\boldsymbol{s}\|_M\le\Delta}} f(\boldsymbol{s}), \tag{3.4}$$

and thus naturally serves as an approximation to the global optimal solution $\boldsymbol{s}_{\text{opt}}$ of (1.2).

Generically, $\dim\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g})$ strictly increases by 1 as k increases by 1 and thus often $\dim\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g}) = k+1$ until $k = n-1$. But it can happen that $\dim\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g})$ may stop increasing at certain k. When that happens, the Lanczos process breaks down and an invariant subspace of $M^{-1}H$ is found. Let $k_{\max}$ be the smallest nonnegative integer such that

$$\dim\mathcal{K}_{k_{\max}}(M^{-1}H, M^{-1}\boldsymbol{g}) = \dim\mathcal{K}_{k_{\max}+1}(M^{-1}H, M^{-1}\boldsymbol{g}) = k_{\max}+1. \tag{3.5}$$

This is reflected by $\gamma_{k_{\max}+1} = 0$ while $\gamma_k \neq 0$ for all $0 \le k \le k_{\max}$. In such a case, $HQ_{k_{\max}} = MQ_{k_{\max}}T_{k_{\max}}$.

Remark 3.1. GLTR can be thought of as an efficient implementation of TLTRS. In GLTR, the (preconditioned) CG is used instead of the (preconditioned) Lanczos process, and in the first pass, the constraint $\|\boldsymbol{s}\|_M \le \Delta$ in (3.4) is skipped implicitly and $\boldsymbol{s}_k$ is updated by CG until $\|\boldsymbol{s}_k\|_M$ exceeds Δ or T_k is detected to be indefinite. GLTR then enters into the second pass, and thereafter $\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g})$ is continuously expanded by (preconditioned) CG and

$$\boldsymbol{s}_k = \arg\min_{\substack{\boldsymbol{s}\in\mathcal{K}_k(M^{-1}H,M^{-1}\boldsymbol{g})\\ \|\boldsymbol{s}\|_M=\Delta}} f(\boldsymbol{s}).$$

Therefore, the second pass of GLTR is mathematically equivalent to TLTRS.

Two special cases for $M = I_n$ are worth mentioning:

1. the case $\boldsymbol{g} = 0$. TLTRS reduces to the classical Lanczos method for finding the smallest eigenpair of H;
2. the case when H is positive definite and $\Delta \ge \|H^{-1}\boldsymbol{g}\|_2$. TLTRS is equivalent to CG for solving the linear system $H\boldsymbol{s} = -\boldsymbol{g}$ (see subsection 4.2 for detail).

In view of these two special cases, we may say that TLTRS or its efficient implementation GLTR lies between the Lanczos-based method for the linear system and that for the symmetric eigenvalue problem.

Remark 3.2. In order to simplify our presentation, in what follows, we assume the weighting matrix $M = I_n$, except in subsection 4.3, and thereby discuss the relations of the optimal values and optimal solutions between the classical problem (1.1) and the projected one (3.3). Mathematically, we will see in subsection 4.3 that doing so does not lose any generality because any convergence result for $M = I_n$ can be translated into one for $M \neq I_n$ through the following substitutions

$$H \leftarrow M^{-1/2} H M^{-1/2}, \quad \boldsymbol{g} \leftarrow M^{-1/2}\boldsymbol{g}.$$

Making $M = I_n$ simplifies Q_k to having orthonormal columns, i.e., $Q_k^\top Q_k = I_{k+1}$ and (3.2) to

$$HQ_k - Q_k T_k = \gamma_{k+1}\boldsymbol{q}_{k+1}\boldsymbol{e}_{k+1}^\top, \quad \gamma_0 = \|\boldsymbol{g}\|_2, \quad Q_k\boldsymbol{e}_1 = \boldsymbol{g}/\gamma_0. \tag{3.6}$$

As we previously assumed, let $k_{\max}$ be the smallest nonnegative integer such that (3.5) holds, i.e., the Lanczos process breaks down at iteration $k_{\max}$ and let $k \leq k_{\max} \leq n$. Let $Q_\perp \in \mathbb{R}^{n\times(n-k_{\max}-1)}$ be any orthogonal complementarity of $Q_{k_{\max}}$ such that $Q := [Q_{k_{\max}}, Q_\perp] \in \mathbb{R}^{n\times n}$ is orthogonal. We have

$$\begin{aligned} Q^\top H Q &= \left[\begin{array}{ccccc|ccc|c} \delta_0 & \gamma_1 & & & & & & & \\ \gamma_1 & \delta_1 & \ddots & & & & & & \\ & \ddots & \ddots & \gamma_k & & & & & \\ & & \gamma_k & \delta_k & & \gamma_{k+1} & & & \\ \hline & & & & \gamma_{k+1} & \delta_{k+1} & \ddots & & \\ & & & & & \ddots & \ddots & \gamma_{k_{\max}} & \\ & & & & & & \gamma_{k_{\max}} & \delta_{k_{\max}} & \\ \hline & & & & & & & & Q_\perp^\top H Q_\perp \end{array}\right] \\ &= \left[\begin{array}{c|c} T_k & \gamma_{k+1}\boldsymbol{e}_{k+1}\boldsymbol{e}_1^\top \\ \hline \gamma_{k+1}\boldsymbol{e}_1\boldsymbol{e}_{k+1}^\top & \widetilde{T}_k \end{array}\right] =: T. \end{aligned} \tag{3.7}$$

Denote the eigenvalues of T_k by $\sigma_i^{(k)}$, often called the *Ritz values*, arranged in the nondecreasing order:

$$\sigma_1^{(k)} \leq \sigma_2^{(k)} \leq \cdots \leq \sigma_{k+1}^{(k)}.$$

Associated with every Lanczos step k before a breakdown is the corresponding TRS (3.3). Let $\boldsymbol{h}_k$ and λ_k be the solution of (3.3) and the Lagrangian multiplier for it, respectively, and set $\boldsymbol{s}_k = Q_k\boldsymbol{h}_k$. With these settings, the following lemma follows.

Lemma 3.1. *We have*

(i) *for any* $k = 0, 1, \ldots, k_{\max}$,

$$\theta_i \leq \sigma_i^{(k)} \leq \theta_{n+i-k-1} \quad \textit{for} \quad i = 1, 2, \ldots, k+1;$$

(ii) *for* $0 \le j \le k \le k_{\max}$,

$$\sigma_i^{(k)} \le \sigma_i^{(j)} \quad \textit{for} \quad i = 1, 2, \ldots, j+1;$$

(iii) *in the nondegenerate case,* $\boldsymbol{s}_{k_{\max}} = \boldsymbol{s}_{\rm opt}$ *and* $\lambda_{k_{\max}} = \lambda_{\rm opt}$.

Proof. The inequalities in items (i) and (ii) are straightforward consequences of the Cauchy interlacing inequalities [22].

Item (iii) for the case $\boldsymbol{g} \not\perp \mathcal{E}_1$ has been proved in [9, Theorem 5.7]. We consider the special scenario:

$$\boldsymbol{g} \perp \mathcal{E}_1 \;\; \text{but} \;\; \|(H - \theta_1 I_n)^\dagger \boldsymbol{g}\|_2 > \Delta.$$

Define $\rho(\lambda) := \|(H + \lambda I_n)^\dagger \boldsymbol{g}\|_2$. Use the eigen-decomposition of H to obtain a secular function $\rho(\lambda)$, from which we know that the condition $\boldsymbol{g} \perp \mathcal{E}_1$ implies that $\rho(\lambda)$ is a continuous and nonincreasing function of $\lambda \in (-\theta_{p+1}, +\infty)$. Also note from $Q^\top \boldsymbol{g} = \gamma_0 \boldsymbol{e}_1$ and (3.7) with $\gamma_{k_{\max}+1} = 0$ that for $\lambda > -\theta_{p+1}$

$$\rho(\lambda) = \|(QTQ^\top + \lambda I_n)^\dagger \boldsymbol{g}\|_2 = \gamma_0 \|(T + \lambda I_n)^\dagger \boldsymbol{e}_1\|_2 = \gamma_0 \|(T_{k_{\max}} + \lambda I_{k_{\max}+1})^\dagger \boldsymbol{e}_1\|_2,$$

implying that $\gamma_0 \|(T_{k_{\max}} + \lambda I_{k_{\max}+1})^\dagger \boldsymbol{e}_1\|_2$ is also a continuous and nonincreasing function of $\lambda > -\theta_{p+1}$. Thus, it follows from

$$\rho(\lambda_{k_{\max}}) = \gamma_0 \|(T_{k_{\max}} + \lambda_{k_{\max}} I_{k_{\max}+1})^\dagger \boldsymbol{e}_1\|_2 \le \Delta < \rho(-\theta_1),$$

that $\lambda_{k_{\max}} > -\theta_1$, i.e., $H + \lambda_{k_{\max}} I_n$ is positive definite. Moreover, by [9, Theorem 5.1], $(H + \lambda_{k_{\max}} I_n)\boldsymbol{s}_{k_{\max}} = -\boldsymbol{g}$, which according to Lemma 2.1 and the uniqueness in the nondegenerate case leads to item (iii). □

Lemma 3.1(iii) says that when a breakdown occurs, TLTRS solves the original problem (1.1) exactly for the nondegenerate case. However, in the degenerate case, the solution $\boldsymbol{s}_{\rm opt}$ is of the form (2.3) with $\tau > 0$. As the Lanczos process starting from $\boldsymbol{g}$ cannot extract any information out of the eigenspace $\mathcal{E}_1$, the approximate solution $\boldsymbol{s}_k = Q_k \boldsymbol{h}_k$ does not contain the component of $\tau \boldsymbol{u}$ for any $\boldsymbol{u} \in \mathcal{E}_1$, even for $k = k_{\max}$. In the other word, the projected problem (3.3) can never deliver a sufficiently close approximate model to the original problem (1.1) for the degenerate case. This is fully discussed in [9, Theorem 5.8] with a restarting strategy to cure this problem. Therefore, in our convergence analysis presented in section 4, we are mainly concerned with the nondegenerate case.

We conclude this section with an important result in [18], which claims that the Lagrangian multipliers λ_k monotonically increases with k.

Lemma 3.2 ([18]). *The sequence* $\{\lambda_k\}_{k=0}^{k_{\max}}$ *of Lagrangian multipliers associated with* (3.3) *is monotonically nondecreasing in* k.

Combining Lemma 3.1 and Lemma 3.2, we have the following Proposition 3.1. This is the first step for the convergence analysis of TLTRS/GLTR.

Proposition 3.1. *Let* $0 \le k \le k_{\max}$, *then*

(i) *if* $\lambda_k = 0$, *then* $\lambda_i = 0$ *for* $i = 0, 1, \ldots, k$, *and*

(ii) *in the nondegenerate case,* $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{k_{\max}} = \lambda_{\rm opt}$.

4 Convergence analysis for TLTRS

Throughout this section, we assume that (1.1) is nondegenerate, unless otherwise explicitly stated differently. Also, since GLTR is an efficient implementation of TLTRS and is essentially equivalent to TLTRS (cf. Remark 3.1), our convergence analysis will focus on TLTRS only.

We will analyze the convergence for the two cases $\lambda_{\rm opt} = 0$ and $\lambda_{\rm opt} > 0$, separately.

Let $\mathbb{P}_k$ denote all polynomials with degree no higher than k. The Chebyshev polynomials will show up in our convergence analysis. The kth Chebyshev polynomial of the first kind $\mathscr{T}_k(t) \in \mathbb{P}_k$ is

$$\begin{aligned}\mathscr{T}_k(t) &= \cos(k \arccos t) && \text{for } |t| \le 1,\\ &= \frac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^k + \left(t + \sqrt{t^2-1}\right)^{-k}\right] && \text{for } |t| \ge 1.\end{aligned}$$

It frequently shows up in numerical analysis and computations because of its numerous nice properties, for example $|\mathscr{T}_k(t)| \le 1$ for $|t| \le 1$ and $|\mathscr{T}_k(t)|$ grows extremely fast[4] for $|t| > 1$. We have the following classical result (see e.g., [28, Theorem 6.25]).

Lemma 4.1. *For a given $\gamma \notin [a, b]$, we have*

$$\min_{p\in\mathbb{P}_k, p(\gamma)=1} \max_{t\in[a,b]} |p(t)| = \left|\mathscr{T}_k\left(1 + 2\frac{\gamma - b}{b - a}\right)\right|^{-1} = \left|\mathscr{T}_k\left(-\frac{[b-\gamma]+[a-\gamma]}{[b-\gamma]-[a-\gamma]}\right)\right|^{-1}. \quad (4.1)$$

There is an elegant expression for $\mathscr{T}_k(\cdots)$ in (4.1), namely [15],

$$\left|\mathscr{T}_k\left(\frac{1+t}{1-t}\right)\right| = \left|\mathscr{T}_k\left(\frac{t+1}{t-1}\right)\right| = \frac{1}{2}\left(\mathit{\Gamma}_t^k + \mathit{\Gamma}_t^{-k}\right) \quad \text{for } 1 \ne t > 0, \quad (4.2a)$$

where

$$\mathit{\Gamma}_t := \frac{\sqrt{t}+1}{|\sqrt{t}-1|} \quad \text{for } t > 0. \quad (4.2b)$$

4.1 Convergence when $\lambda_{\rm opt} = 0$

In this case, H is positive definite, and moreover $\|H^{-1}\boldsymbol{g}\|_2 \le \mathit{\Delta}$, implying that (1.1) is equivalent to the linear system: $H\boldsymbol{s}_{\rm opt} = -\boldsymbol{g}$, and

$$f(\boldsymbol{s}) = \frac{1}{2}(\boldsymbol{s}_{\rm opt} - \boldsymbol{s})^\top H(\boldsymbol{s}_{\rm opt} - \boldsymbol{s}) - \frac{1}{2}\boldsymbol{s}_{\rm opt}^\top H\boldsymbol{s}_{\rm opt}.$$

Furthermore, by Lemma 3.1(iii) and Proposition 3.1, we know that $\lambda_k = 0$ for all $k = 0, 1, \ldots, k_{\max}$, which implies that each TRS (3.3) is equivalent to the linear system: $T_k\boldsymbol{h}_k = -\gamma_0\boldsymbol{e}_1$, and TLTRS turns out to be the full orthogonalization method (FOM) [28, Algorithm 6.4]. Indeed,

$$\boldsymbol{s}_k = Q_k\boldsymbol{h}_k = \arg\min_{\boldsymbol{s}\in\mathcal{K}_k(H,\boldsymbol{g})} \frac{1}{2}(\boldsymbol{s}_{\rm opt} - \boldsymbol{s})^\top H(\boldsymbol{s}_{\rm opt} - \boldsymbol{s}),$$

[4] In fact, a result due to Chebyshev himself says that if $p(t)$ is a polynomial of degree no bigger than k and $|p(t)| \le 1$ for $-1 \le t \le 1$, then $|p(t)| \le |\mathscr{T}_k(t)|$ for any t outside $[-1, 1]$ [3, p.65].

the same as the one obtained from CG [28, Section 6.7] on the linear system $H\boldsymbol{s}_{\rm opt} = -\boldsymbol{g}$. In the other word, in this case, GLTR will never go over the boundary of $\|\boldsymbol{s}\|_2 \le \Delta$ (i.e., the second pass of GLTR will never be called), and thereby, the approximation $\boldsymbol{s}_k$ from GLTR is the same as that from the CG iteration for $H\boldsymbol{s}_{\rm opt} = -\boldsymbol{g}$. Consequently, the standard convergence theory [28, Section 6.11.3] for CG applies for this situation. In particular, we have

$$\frac{\|\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k\|_H}{\|\boldsymbol{s}_{\rm opt}\|_H} \le \frac{1}{\mathscr{T}_{k+1}((\kappa+1)/(\kappa-1))} = \frac{2}{\Gamma_\kappa^{k+1} + \Gamma_\kappa^{-(k+1)}}, \tag{4.3}$$

where Γ_κ is defined by (4.2), and $\kappa := \kappa(H) = \frac{\theta_n}{\theta_1}$ is the spectral condition number of H. In terms of the spectral norm, (4.3) implies

$$\frac{\|\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k\|_2}{\|\boldsymbol{s}_{\rm opt}\|_2} \le \frac{\sqrt{\kappa}}{\mathscr{T}_{k+1}((\kappa+1)/(\kappa-1))},$$

and by $f(\boldsymbol{s}_{\rm opt}) = -\frac{1}{2}\boldsymbol{s}_{\rm opt}^\top H\boldsymbol{s}_{\rm opt}$ in this case, it holds that

$$0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) = \frac{1}{2}(\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k)^\top H(\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k) \le \frac{\|H\|_2\kappa\|\boldsymbol{s}_{\rm opt}\|_2^2}{2\mathscr{T}_{k+1}^2\left((\kappa+1)/(\kappa-1)\right)}.$$

4.2 Convergence when $\lambda_{\rm opt} > 0$

If $\lambda_{\rm opt} > 0$, then $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$. This is the case when H is indefinite or positive definite but $\|H^{-1}\boldsymbol{g}\|_2 > \Delta$.

But we point out that $\lambda_{\rm opt} > 0$ does not imply all $\lambda_k > 0$, as the following simple example demonstrates, where $\lambda_{\rm opt} > 0$ but $\lambda_k = 0$ for some $0 \le k \le k_{\max}$, i.e., $\boldsymbol{h}_k = -\gamma_0 T_k^{-1}\boldsymbol{e}_1$ is the solution to the projected TRS (3.3).

Example 4.1. Consider TRS with

$$H = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{g} = \boldsymbol{e}_1 \in \mathbb{R}^2 \text{ and } \Delta = 1.$$

It can be verified that $\lambda_{\rm opt} \approx 0.1701$ and $\boldsymbol{s}_{\rm opt} \approx [-0.7602, 0.6497]^\top$, but $\lambda_0 = 0$ and $\boldsymbol{s}_0 = -\boldsymbol{e}_1/2$.

Even though $\lambda_k = 0$ may happen in the early stage of TLTRS, eventually $\lambda_k > 0$ as k increases, and thereby $\|\boldsymbol{s}_k\|_2 = \Delta$. This also means that GLTR will eventually encounter the boundary of $\|\boldsymbol{s}\|_2 \le \Delta$, and proceeds to the second pass and the third pass. For that reason, in what follows, we analyze the errors

$$\|\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k\|_2 \quad \text{and} \quad |f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt})|$$

under the assumption $\|\boldsymbol{s}_k\|_2 = \Delta$. Set

$$H_{\rm opt} := H + \lambda_{\rm opt} I_n \tag{4.4}$$

which is positive definite since it is assumed in the nondegenerate case.

Before proceeding further, we mention a related analysis given in [35] of the truncated Conjugate-Gradient (tCG) method for the strictly convex TRS (i.e., H is positive definite).

When CG encounters the boundary, GLTR and tCG invoke different procedures. In particular, tCG stops at the next step by choosing the intersection point of the CG path [35] on the boundary $\|\boldsymbol{s}\|_2 = \Delta$, while GLTR continues from the last CG step by expanding the Krylov subspace and finding a boundary approximation $\boldsymbol{s}_k$ in the Krylov subspace. The main result in [35] shows that the reduction in the objective function by tCG is at least half of the reduction by $\boldsymbol{s}_{\rm opt}$. This result was generalized to the convex case in [4, Section 7.5.2]. There is a major difference in goals between GLTR and tCG: GLTR seeks accurate approximations to $\boldsymbol{s}_{\rm opt}$, as accurate as dictated by chosen tolerance, whereas tCG attempts to find approximations that hopefully reduce the objective function by significant fractions as $\boldsymbol{s}_{\rm opt}$ does. Our analysis in this paper is concerned with GLTR for the general TRS, and will provide upper bounds for the absolute reductions $|f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt})|$ and $\|\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k\|_2$.

4.2.1 The optimal polynomial

Note that the approximate solution $\boldsymbol{s}_k = Q_k \boldsymbol{h}_k \in \mathcal{K}_k(H, \boldsymbol{g})$ can be expressed as

$$\boldsymbol{s}_k = \psi_k(H)\boldsymbol{g} = U\psi_k(\Theta)U^\top \boldsymbol{g} = U\psi_k(\Theta)\boldsymbol{a} = \sum_{i=1}^{n} \psi_k(\theta_i) a_i \boldsymbol{u}_i,$$

where $\boldsymbol{a} = U^\top \boldsymbol{g}$ and the optimal polynomial $\psi_k \in \mathbb{P}_k$ is given by

$$\psi_k = \arg \min_{\psi \in \mathbb{P}_k,\ \|\psi(H)\boldsymbol{g}\|_2 = \Delta} f(\psi(H)\boldsymbol{g}). \tag{4.5}$$

Let

$$\psi_k(\theta) = \sum_{i=0}^{k} \widehat{p}_i \theta^i = \underbrace{[\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_{k+1}]}_{=:\widehat{\boldsymbol{p}}^\top} \begin{bmatrix} 1 \\ \theta \\ \vdots \\ \theta^k \end{bmatrix} \quad \text{and} \quad V_{k+1,n} := \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \theta_1 & \theta_2 & \ldots & \theta_n \\ \vdots & \vdots & \cdots & \vdots \\ \theta_1^k & \theta_2^k & \ldots & \theta_n^k \end{bmatrix}.$$

Since any $\boldsymbol{s} \in \mathcal{K}_k(H, \boldsymbol{g})$ takes the form $\boldsymbol{s} = \psi(H)\boldsymbol{g} = U\psi(\Theta)\boldsymbol{a} = UD_{\boldsymbol{a}}V_{k+1,n}^\top \boldsymbol{p}$, where $D_{\boldsymbol{a}} = \mathrm{diag}(a_1, a_2, \cdots, a_n)$ and $\boldsymbol{p} \in \mathbb{R}^{k+1}$ is the coefficient vector of ψ, we can express the coefficient vector $\widehat{\boldsymbol{p}}$ corresponding to the optimal ψ_k as

$$\widehat{\boldsymbol{p}} = \arg \min_{\|D_{\boldsymbol{a}}V_{k+1,n}^\top \boldsymbol{p}\|_2 = \Delta} \frac{1}{2}\boldsymbol{p}^\top (V_{k+1,n} D_{\boldsymbol{a}} \Theta D_{\boldsymbol{a}} V_{k+1,n}^\top)\boldsymbol{p} + \boldsymbol{p}^\top V_{k+1,n} D_{\boldsymbol{a}} \boldsymbol{a}.$$

By the Lagrangian multiplier theory, there is $\vartheta_k \in \mathbb{R}$ such that

$$\left(V_{k+1,n} D_{\boldsymbol{a}} (\Theta + \vartheta_k I_n) D_{\boldsymbol{a}} V_{k+1,n}^\top\right) \widehat{\boldsymbol{p}} = -V_{k+1,n} D_{\boldsymbol{a}} \boldsymbol{a} \quad \text{and} \quad \|D_{\boldsymbol{a}} V_{k+1,n}^\top \widehat{\boldsymbol{p}}\|_2 = \Delta.$$

Due to the emergence of the Lagrangian multiplier ϑ_k, however, this characterization for the optimal ψ_k doesn't lead to a simple convergence analysis, as opposed to the ones in [14, 15] for analyzing CG and the minimal residual method. In what follows, we adopt an approach of using sub-optimal polynomial approximations to establish bounds on the errors in the approximation solutions.

4.2.2 Solutions resulted from sub-optimal polynomials

Recall that we will be focusing on the situation where $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$ and on approximations $\boldsymbol{s}$ with $\|\boldsymbol{s}\|_2 = \Delta$. We first present a general framework to bound the errors of

$$f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) \quad \text{and} \quad \|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2$$

in terms of any nonzero $\widetilde{\boldsymbol{s}} \in \mathcal{K}_k(H, \boldsymbol{g})$. Later, this framework will be realized for

$$\widetilde{\boldsymbol{s}} = \wp_k(H)\boldsymbol{g} \in \mathbb{R}^n \tag{4.6}$$

constructed from certain sub-optimal polynomial $\wp_k \in \mathbb{P}_k$ as opposed to the optimal one given by (4.5) for the purpose of establishing error bounds for TLTRS solutions (i.e., for the second pass of GLTR).

Theorem 4.1. *Suppose* (1.1) *is nondegenerate,* $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$ *and* $\boldsymbol{s}_k$ *is the* k*th* ($k \le k_{\max}$) *approximation of* TLTRS *satisfying* $\|\boldsymbol{s}_k\|_2 = \Delta$. *Then for any nonzero* $\widetilde{\boldsymbol{s}} \in \mathcal{K}_k(H, \boldsymbol{g})$, *we have*

$$0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) \le 2\|H_{\rm opt}\|_2 \|\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt}\|_2^2, \quad \textit{and} \tag{4.7}$$

$$\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2 \le 2\sqrt{\varkappa}\, \|\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt}\|_2, \tag{4.8}$$

where $H_{\rm opt}$ *is given by* (4.4) *and*

$$\varkappa := \kappa(H_{\rm opt}) = \frac{\theta_n + \lambda_{\rm opt}}{\theta_1 + \lambda_{\rm opt}} \tag{4.9}$$

is the spectral condition number of $H_{\rm opt}$.

Proof. First, we have $\big|\|\widetilde{\boldsymbol{s}}\|_2 - \Delta\big| = \big|\|\widetilde{\boldsymbol{s}}\|_2 - \|\boldsymbol{s}_{\rm opt}\|_2\big| \le \|\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt}\|_2$ which leads to

$$\left|1 - \frac{\Delta}{\|\widetilde{\boldsymbol{s}}\|_2}\right| \le \frac{\|\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt}\|_2}{\|\widetilde{\boldsymbol{s}}\|_2}. \tag{4.10}$$

Let $\boldsymbol{r} = \boldsymbol{v} - \boldsymbol{s}_{\rm opt}$ where $\boldsymbol{v} = (\widetilde{\boldsymbol{s}}/\|\widetilde{\boldsymbol{s}}\|_2)\Delta$. We then have

$$\begin{aligned}
\|\boldsymbol{r}\|_2 = \|\boldsymbol{s}_{\rm opt} - \boldsymbol{v}\|_2 &\le \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2 + \|\widetilde{\boldsymbol{s}} - \boldsymbol{v}\|_2 \\
&= \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2 + \left\|\widetilde{\boldsymbol{s}} - \Delta \cdot \frac{\widetilde{\boldsymbol{s}}}{\|\widetilde{\boldsymbol{s}}\|_2}\right\|_2 \\
&= \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2 + \|\widetilde{\boldsymbol{s}}\|_2 \times \left|1 - \frac{\Delta}{\|\widetilde{\boldsymbol{s}}\|_2}\right| \\
&\le 2\|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2,
\end{aligned} \tag{4.11}$$

where the last inequality is obtained by using (4.10). Moreover, since for any $0 \le i \le k_{\max} - 1$,

$$f(\boldsymbol{s}_i) = \min_{\substack{\boldsymbol{s} \in \mathcal{K}_i(H,\boldsymbol{g}) \\ \|\boldsymbol{s}\|_2 \le \Delta}} f(\boldsymbol{s}) \ge \min_{\substack{\boldsymbol{s} \in \mathcal{K}_{i+1}(H,\boldsymbol{g}) \\ \|\boldsymbol{s}\|_2 \le \Delta}} f(\boldsymbol{s}) = f(\boldsymbol{s}_{i+1}) \ge \min_{\|\boldsymbol{s}\|_2 \le \Delta} f(\boldsymbol{s}) = f(\boldsymbol{s}_{\rm opt}), \tag{4.12}$$

we have

$$\begin{aligned}
0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) &\le f(\boldsymbol{v}) - f(\boldsymbol{s}_{\rm opt}) \\
&= \frac{\boldsymbol{r}^\top H \boldsymbol{r}}{2} + \boldsymbol{r}^\top (H\boldsymbol{s}_{\rm opt} + \boldsymbol{g}) = \frac{\boldsymbol{r}^\top H \boldsymbol{r}}{2} - \lambda_{\rm opt} \boldsymbol{r}^\top \boldsymbol{s}_{\rm opt} \\
&= \frac{\boldsymbol{r}^\top (H + \lambda_{\rm opt} I)\boldsymbol{r}}{2} \qquad (4.13) \\
&\le \frac{\|H_{\rm opt}\|_2}{2} \|\boldsymbol{r}\|_2^2 \\
&\le 2\|H_{\rm opt}\|_2 \, \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2^2, \qquad (4.14)
\end{aligned}$$

where for obtaining (4.13) we have used

$$\Delta^2 = \|\boldsymbol{v}\|_2^2 = \|\boldsymbol{s}_{\rm opt}\|_2^2 + \|\boldsymbol{r}\|_2^2 + 2\,\boldsymbol{r}^\top \boldsymbol{s}_{\rm opt}$$

to get $\boldsymbol{r}^\top \boldsymbol{s}_{\rm opt} = -\|\boldsymbol{r}\|_2^2/2 = -\boldsymbol{r}^\top \boldsymbol{r}/2$, and used (4.11) for getting (4.14). This completes the proof of (4.7).

For (4.8), we define

$$f_{\rm opt}(\boldsymbol{s}) := \frac{1}{2}\boldsymbol{s}^\top H_{\rm opt} \boldsymbol{s} + \boldsymbol{s}^\top \boldsymbol{g} = f(\boldsymbol{s}) + \frac{1}{2}\lambda_{\rm opt}\|\boldsymbol{s}\|_2^2.$$

Then by noting that $\nabla f_{\rm opt}(\boldsymbol{s}_{\rm opt}) = H_{\rm opt}\boldsymbol{s}_{\rm opt} + \boldsymbol{g} = \boldsymbol{0}$, we have for any $\boldsymbol{s}$,

$$f_{\rm opt}(\boldsymbol{s}) = f_{\rm opt}(\boldsymbol{s}_{\rm opt}) + \frac{1}{2}(\boldsymbol{s} - \boldsymbol{s}_{\rm opt})^\top H_{\rm opt}(\boldsymbol{s} - \boldsymbol{s}_{\rm opt}),$$

and thus,

$$f_{\rm opt}(\boldsymbol{s}) - f_{\rm opt}(\boldsymbol{s}_{\rm opt}) \ge \frac{1}{2}(\theta_1 + \lambda_{\rm opt})\|\boldsymbol{s} - \boldsymbol{s}_{\rm opt}\|_2^2. \qquad (4.15)$$

Furthermore, if $\|\boldsymbol{s}\|_2 = \Delta$, then

$$\begin{aligned}
f_{\rm opt}(\boldsymbol{s}) - f_{\rm opt}(\boldsymbol{s}_{\rm opt}) &= \big[f(\boldsymbol{s}) + \frac{1}{2}\lambda_{\rm opt}\|\boldsymbol{s}\|_2^2\big] - \big[f(\boldsymbol{s}_{\rm opt}) + \frac{1}{2}\lambda_{\rm opt}\|\boldsymbol{s}_{\rm opt}\|_2^2\big] \\
&= f(\boldsymbol{s}) - f(\boldsymbol{s}_{\rm opt})
\end{aligned}$$

since $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$ also. Consequently, for $\boldsymbol{s}_k$, by (4.7) and (4.15), we have

$$\frac{1}{2}(\theta_1 + \lambda_{\rm opt})\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2^2 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) \le 2\|H_{\rm opt}\|_2 \, \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2^2,$$

yielding (4.8). □

Next, we will discuss two sub-optimal polynomials $\wp_k \in \mathbb{P}_k$ to realize $\widetilde{\boldsymbol{s}}$ by (4.6).

Polynomials resulting from $\min \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2$**.** According to Theorem 4.1, a good bound for $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2$ can be pursued by minimizing $\|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2$ over $\widetilde{\boldsymbol{s}} \in \mathcal{K}_k(H, \boldsymbol{g})$. Note that $\widetilde{\boldsymbol{s}} \in \mathcal{K}_k(H, \boldsymbol{g}) = \mathcal{K}_k(H_{\rm opt}, \boldsymbol{g})$, and therefore,

$$\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt} = \wp_k(H_{\rm opt})\boldsymbol{g} + H_{\rm opt}^{-1}\boldsymbol{g}$$

$$
\begin{aligned}
&= (\wp_k(H_{\rm opt})H_{\rm opt} + I_n)H_{\rm opt}^{-1}\boldsymbol{g} \\
&= -\hbar(H_{\rm opt})\boldsymbol{s}_{\rm opt},
\end{aligned}
$$

where $\hbar(t) := 1 + t\wp_k(t) \in \mathbb{P}_{k+1}$ satisfying $\hbar(0) = 1$. Hence, noting $\|\boldsymbol{s}_{\rm opt}\|_2 \le \Delta$, we have

$$
\begin{aligned}
\min_{\wp\in\mathbb{P}_k} \|\widetilde{\boldsymbol{s}} - \boldsymbol{s}_{\rm opt}\|_2 &\le \min_{\hbar\in\mathbb{P}_{k+1},\hbar(0)=1} \|\hbar(H_{\rm opt})\|_2 \cdot \Delta \\
&\le \min_{\hbar\in\mathbb{P}_{k+1},\hbar(0)=1} \max_{i=1,2,\ldots,n} |\hbar(\theta_i + \lambda_{\rm opt})| \cdot \Delta \\
&\le \min_{\hbar\in\mathbb{P}_{k+1},\hbar(0)=1} \max_{t\in[\theta_1+\lambda_{\rm opt},\theta_n+\lambda_{\rm opt}]} |\hbar(t)| \cdot \Delta \\
&= \frac{\Delta}{\mathscr{T}_{k+1}(\eta)} = 2\Delta\left(\Gamma_{\varkappa}^{k+1} + \Gamma_{\varkappa}^{-(k+1)}\right)^{-1}, \qquad (4.16)
\end{aligned}
$$

where the second equality in (4.16) follows from Lemma 4.1 and (4.2), and

$$
\eta := \frac{\varkappa+1}{\varkappa-1} = 1 + 2\frac{\theta_1 + \lambda_{\rm opt}}{\theta_n - \theta_1} \qquad (4.17)
$$

with $\varkappa = \kappa(H_{\rm opt})$ defined by (4.9). Substituting (4.16) into (4.7) and (4.8) gives

$$
0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) \le 2\|H_{\rm opt}\|_2 \left(\frac{\Delta}{\mathscr{T}_{k+1}(\eta)}\right)^2, \qquad (4.18a)
$$

$$
\|\boldsymbol{s}_{\rm opt} - \boldsymbol{s}_k\|_2 \le \frac{2\sqrt{\varkappa}\Delta}{\mathscr{T}_{k+1}(\eta)}. \qquad (4.18b)
$$

Best polynomials for approximating $\frac{1}{x-\eta}$. We next discuss yet another sub-optimal polynomial. Note that $\boldsymbol{g} = U\boldsymbol{a} = \sum_{i=1}^n \boldsymbol{u}_i a_i$, and we have for any $\boldsymbol{s} \in \mathcal{K}_k(H, \boldsymbol{g})$

$$
\boldsymbol{s} = \psi(H)\boldsymbol{g} = U\psi(\Theta)\boldsymbol{a} = \sum_{i=1}^n \psi(\theta_i)a_i\boldsymbol{u}_i \quad \text{for some } \psi \in \mathbb{P}_k, \text{ and} \qquad (4.19)
$$

$$
\boldsymbol{s}_{\rm opt} = -(H + \lambda_{\rm opt}I_n)^{-1}\boldsymbol{g} = -U(\Theta + \lambda_{\rm opt}I_n)^{-1}\boldsymbol{a} = -\sum_{i=1}^n \frac{a_i}{\theta_i + \lambda_{\rm opt}}\boldsymbol{u}_i. \qquad (4.20)
$$

By comparing (4.19) and (4.20), we define a sub-optimal polynomial $\wp_k^{\rm ra} \in \mathbb{P}_k$ as the solution to the following minimax approximation problem:

$$
\wp_k^{\rm ra} := \arg\min_{\wp\in\mathbb{P}_k} \max_{\theta_1\le\theta\le\theta_n} \left|\wp(\theta) - \frac{1}{\theta + \lambda_{\rm opt}}\right|. \qquad (4.21)
$$

In the other word, $\wp_k^{\rm ra}$ is the *best polynomial of approximation* to the rational function $\frac{1}{\theta+\lambda_{\rm opt}}$ in the interval $[\theta_1, \theta_n]$.

Note that the linear transformation

$$
\theta(x) = \frac{\theta_n - \theta_1}{2}x + \frac{\theta_1 + \theta_n}{2}
$$

maps $x \in [-1, 1]$ one-to-one and onto $\theta \in [\theta_1, \theta_n]$; moreover, by (4.17), $\eta > 1$, and we have

$$
\min_{\psi\in\mathbb{P}_k} \max_{\theta_1\le\theta\le\theta_n} \left|\psi(\theta) - \frac{1}{\theta + \lambda_{\rm opt}}\right|
$$

$$
\begin{aligned}
&= \min_{\psi\in\mathbb{P}_k} \max_{-1\le x\le 1} \left| \psi(\theta(x)) - \frac{2}{(\theta_n-\theta_1)(x-\frac{\theta_1+\theta_n+2\lambda_{\rm opt}}{\theta_n-\theta_1})} \right| \\
&= \frac{2}{\theta_n-\theta_1} \times \min_{\psi\in\mathbb{P}_k} \max_{-1\le x\le 1} \left| \frac{(\theta_n-\theta_1)\psi(\theta(x))}{2} - \frac{1}{x-\frac{\theta_1+\theta_n+2\lambda_{\rm opt}}{\theta_n-\theta_1}} \right| \\
&= \frac{2}{\theta_n-\theta_1} \times \min_{\wp\in\mathbb{P}_k} \max_{-1\le x\le 1} \left| \wp(x) - \frac{1}{x-\eta} \right|, \qquad \left(\text{with } \wp(x) = \tfrac{(\theta_n-\theta_1)\psi(\theta(x))}{2}\right)
\end{aligned}
$$

which implies that

$$
\max_{\theta_1\le\theta\le\theta_n} \left| \wp_k^{\rm ra}(\theta) - \frac{1}{\theta+\lambda_{\rm opt}} \right| = \frac{2}{\theta_n-\theta_1} \times \underbrace{\min_{\wp\in\mathbb{P}_k} \max_{-1\le x\le 1} \left| \wp(x) - \frac{1}{x-\eta} \right|}_{=:\epsilon_k^{\rm ra}(\eta)}, \tag{4.22}
$$

where $\epsilon_k^{\rm ra}(\eta)$ is the error of approximation by the best polynomial in $\mathbb{P}_k$ to $\frac{1}{x-\eta}$ in the interval $[-1,1]$.

For the behavior of $\epsilon_k^{\rm ra}(\eta)$ with respect to k and η, we fortunately have the explicit formulation by the pioneering works of Chebyshev and Bernstein. Indeed, Chebyshev found an explicit expression for the best approximating polynomial of $\frac{1}{x-\eta}$ in $[-1,1]$, and Bernstein gave a trigonometric representation as stated in the next lemma [19, Section 4.3].

Lemma 4.2 (Bernstein [2])**.** *Given $\eta>1$, the best approximating polynomial $p_k(x)\in\mathbb{P}_k$ of $\frac{1}{x-\eta}$ in $[-1,1]$ satisfies*

$$
\frac{1}{x-\eta} - p_k(x) = \frac{\left(\eta+\sqrt{\eta^2-1}\,\right)^{-k}}{\eta^2-1}\cos(k\alpha+\beta),
$$

where α and β are such that $x=\cos\alpha$ and $\frac{\eta x-1}{x-\eta}=\cos\beta$, and moreover,

$$
\epsilon_k^{\rm ra}(\eta) := \min_{\wp\in\mathbb{P}_k} \max_{-1\le x\le 1} \left| \wp(x) - \frac{1}{x-\eta} \right| = \frac{\left(\eta+\sqrt{\eta^2-1}\,\right)^{-k}}{\eta^2-1}. \tag{4.23}
$$

Remark 4.1. It is noted that $\eta+\sqrt{\eta^2-1}>\eta>1$ since $\eta>1$, and for η given by (4.17),

$$
\eta+\sqrt{\eta^2-1} = \frac{\sqrt{\varkappa}+1}{\sqrt{\varkappa}-1} = \Gamma_\varkappa,
$$

where $\Gamma_\varkappa$ is defined by (4.2b). Therefore, $\epsilon_k^{\rm ra}(\eta)$ converges linearly to zero with the linear factor $\Gamma_\varkappa^{-1}$ as k increases.

Now we can establish error bounds for TLTRS solutions in terms of $\epsilon_k^{\rm ra}(\eta)$. The corresponding estimates for $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ and $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ also reflect the behavior of TLTRS characterized by the number of Lanczos step k and the parameter η as we will see from the numerical examples in section 5.

Let $\wp_k^{\rm ra}$ be defined by (4.21), and set $\widetilde{\boldsymbol{s}}_k^{\rm ra} := \wp_k^{\rm ra}(H)\boldsymbol{g}$. Note by (4.22) that

$$
\|\widetilde{\boldsymbol{s}}_k^{\rm ra} - \boldsymbol{s}_{\rm opt}\|_2 = \left\| U\left(\wp_k^{\rm ra}(\Theta) - (\Theta+\lambda_{\rm opt})^{-1}\right)\boldsymbol{a} \right\|_2
$$

$$
\begin{aligned}
&\le \|\boldsymbol{g}\|_2 \times \max_{\theta\in\{\theta_1\ldots,\theta_n\}} \left| \wp_k^{\mathrm{ra}}(\theta) - \frac{1}{\theta+\lambda_{\mathrm{opt}}} \right| \\
&\le \|\boldsymbol{g}\|_2 \times \max_{\theta_1\le\theta\le\theta_n} \left| \wp_k^{\mathrm{ra}}(\theta) - \frac{1}{\theta+\lambda_{\mathrm{opt}}} \right| \\
&= \frac{2\|\boldsymbol{g}\|_2}{\theta_n-\theta_1}\epsilon_k^{\mathrm{ra}}(\eta).
\end{aligned}
$$

Now with $\widetilde{\boldsymbol{s}} = \widetilde{\boldsymbol{s}}_k^{\mathrm{ra}}$, by Theorem 4.1, we have

$$
0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\mathrm{opt}}) \le 2\|H_{\mathrm{opt}}\|_2 \left(\frac{2\|\boldsymbol{g}\|_2}{\theta_n-\theta_1}\epsilon_k^{\mathrm{ra}}(\eta) \right)^2, \tag{4.24a}
$$

$$
\|\boldsymbol{s}_{\mathrm{opt}} - \boldsymbol{s}_k\|_2 \le 2\sqrt{\varkappa}\frac{2\|\boldsymbol{g}\|_2}{\theta_n-\theta_1}\epsilon_k^{\mathrm{ra}}(\eta) = 4\frac{\sqrt{\varkappa}\|\boldsymbol{g}\|_2}{\theta_n-\theta_1}\epsilon_k^{\mathrm{ra}}(\eta). \tag{4.24b}
$$

Summarizing the results in (4.18) and (4.24) for the two sub-optimal solutions yields the item (ii) of the following theorem.

Theorem 4.2. *Let $\theta_1 \le \theta_2 \le \cdots \le \theta_n$ be the ordered eigenvalues of H and λ_{opt} be the optimal Lagrange multiplier as in* Theorem 2.1 *for* TRS (1.1)*. Let the sequence $\{\boldsymbol{s}_k\}_{k=0}^{k_{\max}}$ be generated by* TLTRS *for* (1.1)*.*

(i) *The sequence $\{f(\boldsymbol{s}_k)\}_{k=0}^{k_{\max}}$ is nonincreasing, and $f(\boldsymbol{s}_{k_{\max}}) = f(\boldsymbol{s}_{\mathrm{opt}})$ for the* nondegenerate *case, and*

$$
f(\boldsymbol{s}_{k_{\max}}) + \frac{\tau^2\theta_1}{2} \le f(\boldsymbol{s}_{\mathrm{opt}}) \le f(\boldsymbol{s}_{k_{\max}}) \tag{4.25}
$$

for the degenerate *case, where $\tau^2 = \Delta^2 - \|(H-\theta_1 I_n)^\dagger \boldsymbol{g}\|_2^2 \ge 0$ and $\theta_1 \le 0$.*

(ii) *For the* nondegenerate *case, if $\|\boldsymbol{s}_{\mathrm{opt}}\|_2 = \|\boldsymbol{s}_k\|_2 = \Delta$ for some $0 \le k \le k_{\max}$, then*

$$
0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\mathrm{opt}}) \le 2\|H_{\mathrm{opt}}\|_2\zeta_k^2, \tag{4.26a}
$$

$$
\|\boldsymbol{s}_{\mathrm{opt}} - \boldsymbol{s}_k\|_2 \le 2\sqrt{\varkappa}\,\zeta_k, \tag{4.26b}
$$

where $H_{\mathrm{opt}} = H + \lambda_{\mathrm{opt}} I_n$, $\varkappa = \kappa(H_{\mathrm{opt}})$ by (4.9)*, and*

$$
\zeta_k = \min\left\{ \frac{\Delta}{\mathscr{T}_{k+1}(\eta)}, \frac{2\|\boldsymbol{g}\|_2\epsilon_k^{\mathrm{ra}}(\eta)}{\theta_n-\theta_1} \right\}, \tag{4.27}
$$

$\epsilon_k^{\mathrm{ra}}(\eta)$ is defined by (4.23)*, and $\mathscr{T}_{k+1}(\eta)$ is the $(k+1)$st Chebyshev polynomial of the first kind evaluated at η given in* (4.17)*.*

Proof. Based on our previous discussions, only the inequality (4.25) needs a proof. First, it can be seen that $f(\boldsymbol{s}_{k_{\max}})$ is an upper bound for $f(\boldsymbol{s}_{\mathrm{opt}})$ by (4.12). For the orthogonal matrix $Q := [Q_{k_{\max}}, Q_\perp] \in \mathbb{R}^{n\times n}$ satisfying (3.7), let

$$
Q^\top \boldsymbol{s}_{\mathrm{opt}} = \begin{bmatrix} Q_{k_{\max}}^\top \boldsymbol{s}_{\mathrm{opt}} \\ Q_\perp^\top \boldsymbol{s}_{\mathrm{opt}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}.
$$

From $Q^\top H Q = T = \mathrm{diag}(T_{k_{\max}}, T_\perp)$, $T_{k_{\max}} = T_{(1:k_{\max}+1,1:k_{\max}+1)}$, $T_\perp = T_{(k_{\max}+2:n,k_{\max}+2:n)}$ and $Q^\top \boldsymbol{g} = \gamma_0 \boldsymbol{e}_1$, we have

$$
f(\boldsymbol{s}_{\mathrm{opt}}) = \frac{1}{2}\boldsymbol{s}_{\mathrm{opt}}^\top H \boldsymbol{s}_{\mathrm{opt}} + \boldsymbol{g}^\top \boldsymbol{s}_{\mathrm{opt}}
$$

$$
\begin{aligned}
&= \frac{1}{2}\boldsymbol{s}_{\rm opt}^\top QTQ^\top \boldsymbol{s}_{\rm opt} + \boldsymbol{s}_{\rm opt}^\top QQ^\top \boldsymbol{g} \\
&= \frac{1}{2}\boldsymbol{y}^\top T_{k_{\max}}\boldsymbol{y} + \gamma_0 \boldsymbol{y}^\top \boldsymbol{e}_1 + \frac{1}{2}\boldsymbol{z}^\top T_\perp \boldsymbol{z} \\
&\geq f(\boldsymbol{s}_{k_{\max}}) + \frac{1}{2}\boldsymbol{z}^\top T_\perp \boldsymbol{z},
\end{aligned}
$$

where the last inequality holds because $\|\boldsymbol{y}\|_2 = \sqrt{\Delta^2 - \|\boldsymbol{z}\|_2^2} \leq \Delta$ and

$$
\frac{1}{2}\boldsymbol{y}^\top T_{k_{\max}}\boldsymbol{y} + \gamma_0 \boldsymbol{y}^\top \boldsymbol{e}_1 \geq \min_{\|\boldsymbol{h}\|_2 \leq \Delta} f(Q_{k_{\max}}\boldsymbol{h}) = f(\boldsymbol{s}_{k_{\max}}).
$$

We next establish a lower bound for $\frac{1}{2}\boldsymbol{z}^\top T_\perp \boldsymbol{z}$. In fact, we have

$$
\frac{1}{2}\boldsymbol{z}^\top T_\perp \boldsymbol{z} \geq \frac{\theta_1 \|\boldsymbol{z}\|_2^2}{2},
$$

since the condition $\boldsymbol{g} \perp \mathcal{E}_1$ for the degenerate case and the breakdown in the Lanczos process imply $\lambda_{\min}(T_\perp) = \theta_1$.

In what follows, we show $\|\boldsymbol{z}\|_2 = \tau$. To this end, we first note that in the degenerate case

$$
\mathcal{E}_1 = \mathcal{R}(U_1) \subseteq \mathcal{R}(Q_\perp) \quad \text{and} \quad \mathcal{R}(Q_{k_{\max}}) \subseteq \mathcal{R}(U_2), \tag{4.28}
$$

where $U = [U_1, U_2]$ is as in (2.1). It can been seen that

$$
(H - \theta_1 I_n)(H - \theta_1 I_n)^\dagger = U_2 U_2^\top.
$$

By (3.6), we have $(H - \theta_1 I_n)Q_{k_{\max}} = Q_{k_{\max}}(T_{k_{\max}} - \theta_1 I_{k_{\max}+1})$. Pre-multiply both sides by $Q_\perp^\top (H - \theta_1 I_n)^\dagger$ to get

$$
Q_\perp^\top U_2 U_2^\top Q_{k_{\max}} = Q_\perp^\top (H - \theta_1 I_n)^\dagger Q_{k_{\max}}(T_{k_{\max}} - \theta_1 I_{k_{\max}+1}). \tag{4.29}
$$

Since $Q_{k_{\max}}^\top Q_\perp = 0$ and $Q_{k_{\max}}^\top U_1 = 0$ by (4.28), we know that

$$
Q_\perp^\top U_2 U_2^\top Q_{k_{\max}} = Q_\perp^\top (I_n - U_1 U_1^\top) Q_{k_{\max}} = Q_\perp^\top Q_{k_{\max}} - Q_\perp^\top U_1 U_1^\top Q_{k_{\max}} = 0
$$

which, together with (4.29), lead to

$$
Q_\perp^\top (H - \theta_1 I_n)^\dagger Q_{k_{\max}}(T_{k_{\max}} - \theta_1 I_{k_{\max}+1}) = 0. \tag{4.30}
$$

Note $T_{k_{\max}} - \theta_1 I_{k_{\max}+1}$ is positive definite. Post-multiplying both sides of (4.30) by $(T_{k_{\max}} - \theta_1 I_{k_{\max}+1})^{-1}\boldsymbol{e}_1$ and using $Q_{k_{\max}}\boldsymbol{e}_1 = \boldsymbol{g}/\gamma_0$, we get $Q_\perp^\top (H - \theta_1 I_n)^\dagger \boldsymbol{g} = \boldsymbol{0}$, from which and (2.3) it follows that

$$
\boldsymbol{z} = -Q_\perp^\top (H - \theta_1 I_n)^\dagger \boldsymbol{g} + \tau Q_\perp^\top \boldsymbol{u} = \tau Q_\perp^\top \boldsymbol{u} \quad \Rightarrow \quad \|\boldsymbol{z}\|_2 = \tau,
$$

as expected, where $\boldsymbol{u} \in \mathcal{E}_1$ is a unit vector. □

It is noted that in the two sub-optimal solutions, the factor η by (4.17) plays a vital role and, therefore it can serve as some kind of *condition number* for TRS (1.1). In particular, from Theorem 4.2, Lemma 4.2 and Remark 4.1, we observe that both $\frac{1}{\mathscr{T}_{k+1}(\eta)}$ and $\epsilon_k(\eta)^{\rm ra}$ decay fast for big η. Theoretically, we have the following lower bound for η:

Theorem 4.3. *If $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$, then we have*

$$\begin{aligned}\eta &\geq 1 + 2\frac{\|\boldsymbol{g}\|_2 \cos\angle(\boldsymbol{g}, \mathcal{E}_1)}{\sqrt{\Delta^2(\theta_n + \|H\|_2 + \|\boldsymbol{g}\|_2/\Delta)^2 - \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)}} \\ &= 1 + 2\frac{\|\boldsymbol{g}\|_2 \cos\angle(\boldsymbol{g}, \mathcal{E}_1)}{\sqrt{\Delta^2(\theta_n + \|H\|_2)^2 + 2\Delta(\theta_n + \|H\|_2)\|\boldsymbol{g}\|_2 + \|\boldsymbol{g}\|_2^2 \sin^2\angle(\boldsymbol{g}, \mathcal{E}_2)}}, \end{aligned} \tag{4.31}$$

where $\angle(\boldsymbol{g}, \mathcal{E}_1)$ and $\angle(\boldsymbol{g}, \mathcal{E}_2)$ stand for the angles from $\mathcal{R}(\boldsymbol{g})$ to $\mathcal{E}_1 = \mathcal{R}(U_1)$ and $\mathcal{E}_2 = \mathcal{R}(U_2)$, respectively, defined by $\cos\angle(\boldsymbol{g}, \mathcal{E}_i) = \frac{\|U_i^\top \boldsymbol{g}\|_2}{\|\boldsymbol{g}\|_2}$, and U_i for $i = 1, 2$ are defined in the paragraph after Theorem 2.1.

Proof. The assertion is true for the degenerate case, i.e., $\lambda_{\rm opt} = -\theta_1$ and $\eta = 1$, since $\cos\angle(\boldsymbol{g}, \mathcal{E}_1) = 0$.

Now consider the nondegenerate case. Then $\lambda_{\rm opt} > -\theta_1$ and $H_{\rm opt} = H + \lambda_{\rm opt} I_n$ is positive definite. Thus $\max\{0, -\theta_1\} \leq \lambda_{\rm opt}$. For this case, we also have $(H + \lambda_{\rm opt} I_n)\boldsymbol{s}_{\rm opt} = -\boldsymbol{g}$ and $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$. Therefore

$$\lambda_{\rm opt}\Delta = \|\lambda_{\rm opt}\boldsymbol{s}_{\rm opt}\|_2 = \| - \boldsymbol{g} - H\boldsymbol{s}_{\rm opt}\|_2 \leq \|\boldsymbol{g}\|_2 + \|H\|_2\Delta.$$

Putting all together, we have

$$\max\{0, -\theta_1\} \leq \lambda_{\rm opt} \leq \|H\|_2 + \frac{\|\boldsymbol{g}\|_2}{\Delta}. \tag{4.32}$$

Also, it follows from $\|\boldsymbol{s}_{\rm opt}\|_2 = \Delta$ and (4.20) that

$$\begin{aligned}\Delta^2 = \sum_{i=1}^{n} \frac{(\boldsymbol{u}_i^\top \boldsymbol{g})^2}{(\theta_i + \lambda_{\rm opt})^2} &= \sum_{i=1}^{p} \frac{(\boldsymbol{u}_i^\top \boldsymbol{g})^2}{(\theta_1 + \lambda_{\rm opt})^2} + \sum_{i=p+1}^{n} \frac{(\boldsymbol{u}_i^\top \boldsymbol{g})^2}{(\theta_i + \lambda_{\rm opt})^2} \\ &\geq \left(\frac{\cos^2\angle(\boldsymbol{g}, \mathcal{E}_1)}{(\theta_1 + \lambda_{\rm opt})^2} + \frac{\cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)}{(\theta_n + \lambda_{\rm opt})^2}\right)\|\boldsymbol{g}\|_2^2.\end{aligned}$$

Multiply both sides by $(\theta_1 + \lambda_{\rm opt})^2(\theta_n + \lambda_{\rm opt})^2$ to get

$$\begin{aligned}(\theta_1 + \lambda_{\rm opt})^2\left[(\theta_n + \lambda_{\rm opt})^2\Delta^2 - \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)\right] &\geq \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_1)(\theta_n + \lambda_{\rm opt})^2 \\ &\geq \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_1)(\theta_n - \theta_1)^2,\end{aligned}$$

where we have used $\lambda_{\rm opt} > -\theta_1$ for obtaining the last inequality. Therefore,

$$\begin{aligned}\frac{\theta_1 + \lambda_{\rm opt}}{\theta_n - \theta_1} &\geq \frac{\|\boldsymbol{g}\|_2 \cos\angle(\boldsymbol{g}, \mathcal{E}_1)}{\sqrt{(\theta_n + \lambda_{\rm opt})^2\Delta^2 - \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)}} \\ &\geq \frac{\|\boldsymbol{g}\|_2 \cos\angle(\boldsymbol{g}, \mathcal{E}_1)}{\sqrt{\Delta^2(\theta_n + \|H\|_2 + \|\boldsymbol{g}\|_2/\Delta)^2 - \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)}},\end{aligned}$$

and, by (4.17), we have

$$\eta \geq 1 + 2\frac{\|\boldsymbol{g}\|_2 \cos\angle(\boldsymbol{g}, \mathcal{E}_1)}{\sqrt{\Delta^2(\theta_n + \|H\|_2 + \|\boldsymbol{g}\|_2/\Delta)^2 - \|\boldsymbol{g}\|_2^2 \cos^2\angle(\boldsymbol{g}, \mathcal{E}_2)}}.$$

The quantity within the square root sign in the above can be expressed differently to show that it is always positive:

$$\Delta^2(\theta_n + \|H\|_2)^2 + 2\Delta(\theta_n + \|H\|_2)\|\boldsymbol{g}\|_2 + \|\boldsymbol{g}\|_2^2 \sin^2 \angle(\boldsymbol{g}, \mathcal{E}_2) > 0,$$

and thus (4.31) follows. □

Remark 4.2. As revealed by Theorem 4.2, large η translates into fast convergence of TLTRS (at the second pass of GLTR). There are several factors that influence the value η, and particularly we have

$$\varkappa \approx 1 \xRightarrow{\text{by (4.2)}} \Gamma_\varkappa \text{ is large } \xLeftrightarrow{\text{by } \Gamma_\varkappa = \eta + \sqrt{\eta^2 - 1}} \eta \text{ is large } \xLeftarrow{\text{by (4.31)}} \Delta \text{ is large}, \tag{4.33}$$

other things being equal. This relation (4.33) is well reflected in our numerical examples in section 5.

Remark 4.3. The quantities θ_1, θ_n, λ_{opt}, $\varkappa$ and η involved in the upper bounds in (4.26a), (4.27), and (4.26b) are usually unknown. Fortunately, TLTRS is able to produce approximations to all these quantities: $\sigma_1^{(k)} \approx \theta_1$, $\sigma_{k+1}^{(k)} \approx \theta_n$, $\lambda_k \approx \lambda_{\text{opt}}$,

$$\varkappa = \frac{\theta_n + \lambda_{\text{opt}}}{\theta_1 + \lambda_{\text{opt}}} \approx \frac{\sigma_{k+1}^{(k)} + \lambda_k}{\sigma_1^{(k)} + \lambda_k} =: \varkappa_k \quad \text{and} \quad \eta \approx 1 + 2\frac{\sigma_1^{(k)} + \lambda_k}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}} =: \eta_k, \tag{4.34}$$

and for modest k (usually a couple of tens suffices), these approximations are usually very good, where $\sigma_{k+1}^{(k)}$ and $\sigma_1^{(k)}$ are the largest and smallest eigenvalues of $T_k \in \mathbb{R}^{(k+1)\times(k+1)}$ in (3.1). Note that all the extra computation effort for these approximations is to compute $\sigma_{k+1}^{(k)}$ and $\sigma_1^{(k)}$. As $k \ll n$ in general, obtaining these approximations can be economical (with $O(k^2)$ flops). Thus, practical estimates for $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$, instead of (4.26), are

$$0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}}) \lesssim 2(\sigma_{k+1}^{(k)} + \lambda_k)\chi_k^2 \quad \text{and} \quad \|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2 \lesssim 2\sqrt{\varkappa_k}\,\chi_k, \tag{4.35}$$

where

$$\chi_k = \min\left\{ \frac{\Delta}{\mathscr{T}_{k+1}(\eta_k)}, \frac{2\|\boldsymbol{g}\|_2 \epsilon_k^{\text{ra}}(\eta_k)}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}} \right\}.$$

We will see in our numerical examples in section 5 that these approximations can provide good enough upper bound estimates for $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$.

The accuracy of the estimates (4.35) for the upper bounds in (4.26) depends on the key quantity η_k. In the following, we provide lower and upper bounds for $\eta - \eta_k$. For this purpose, we let $\boldsymbol{w} \in \mathbb{R}^{k+1}$ and $\boldsymbol{v} \in \mathbb{R}^{k+1}$ be the unit eigenvector of T_k associated with $\sigma_1^{(k)}$ and $\sigma_{k+1}^{(k)}$, respectively, and define

$$\omega_k := |\gamma_{k+1}\boldsymbol{w}^\top \boldsymbol{e}_{k+1}| \quad \text{and} \quad \nu_k := |\gamma_{k+1}\boldsymbol{v}^\top \boldsymbol{e}_{k+1}|, \tag{4.36}$$

where γ_{k+1} is given in (3.2). Then we have the following result.

Proposition 4.1. *Under the assumptions in* Theorem 4.2(ii) *and in exact arithmetic, the estimate* η_k *of* η *given in* (4.34) *satisfies*

$$\eta - \eta_k \leq 2\frac{\lambda_{\rm opt} - \lambda_k}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}}. \tag{4.37}$$

Furthermore, if

$$|\theta_n - \sigma_{k+1}^{(k)}| = \min_i |\theta_i - \sigma_{k+1}^{(k)}|, \tag{4.38}$$

then

$$-2\frac{\omega_k(\sigma_{k+1}^{(k)} + \lambda_k) + \nu_k(\sigma_1^{(k)} + \lambda_k)}{\left(\sigma_{k+1}^{(k)} - \sigma_1^{(k)}\right)^2} \leq \eta - \eta_k \leq 2\frac{\lambda_{\rm opt} - \lambda_k}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}}. \tag{4.39}$$

Proof. For (4.37), by Lemma 3.1, it holds that $\sigma_{k+1}^{(k)} \leq \theta_n$ and $\sigma_1^{(k)} \geq \theta_1$, which using (4.17) and (3.1), gives

$$\eta = 1 + 2\frac{\theta_1 + \lambda_{\rm opt}}{\theta_n - \theta_1} \leq 1 + 2\frac{\sigma_1^{(k)} + \lambda_{\rm opt}}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}} = \eta_k + 2\frac{\lambda_{\rm opt} - \lambda_k}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}}.$$

For the left inequality of (4.39), according to [37], in exact arithmetic, the condition (4.38) implies that

$$\sigma_{k+1}^{(k)} \leq \theta_n \leq \sigma_{k+1}^{(k)} + \nu_k \quad \text{and} \quad \sigma_1^{(k)} \geq \theta_1 \geq \sigma_1^{(k)} - \omega_k, \tag{4.40}$$

which, using $\lambda_{\rm opt} \geq \lambda_k$, $\omega_k \geq 0$ and $\nu_k \geq 0$, leads to the left inequality of (4.39). □

We notice from (4.39) that when $\eta_k \leq \eta$, then the accuracy of η_k is dependent on the accuracy of the Lagrange multiplier λ_k, but when $\eta_k \geq \eta$, ω_k and ν_k play a role. For the latter case, the following known results (see, e.g., [5, Theorem 7.2])

$$\|HQ_k\boldsymbol{w} - \sigma_1^{(k)}Q_k\boldsymbol{w}\|_2 = \omega_k \quad \text{and} \quad \|HQ_k\boldsymbol{v} - \sigma_{k+1}^{(k)}Q_k\boldsymbol{v}\|_2 = \nu_k$$

directly quantify the accuracy of the Ritz values $\sigma_1^{(k)}$ and $\sigma_{k+1}^{(k)}$ as approximations to the extreme eigenvalues θ_1 and θ_n, respectively. Thus, (4.39) connects the accuracy of these Ritz values with the accuracy of η_k. Usually the convergence of $\sigma_1^{(k)}$ and $\sigma_{k+1}^{(k)}$ is very rapid, as supported by existing convergence theory of the Lanczos method [17, 22, 27].

4.3 Convergence for the weighted-norm TRS

We next briefly show that our convergence results for $M = I_n$ can be straightforwardly extended to TLTRS (with the preconditioned Lanczos process) for solving the weighted-norm TRS (1.2). Indeed, by defining

$$\boldsymbol{c} := M^{1/2}\boldsymbol{s}, \quad \widehat{H} := M^{-1/2}HM^{-1/2}, \quad \text{and} \quad \widehat{\boldsymbol{g}} := M^{-1/2}\boldsymbol{g},$$

we can rewrite the weighted-norm TRS (1.2) equivalently as

$$\min_{\|\boldsymbol{c}\|_2 \leq \Delta} \widehat{f}(\boldsymbol{c}) \quad \text{with} \quad \widehat{f}(\boldsymbol{c}) := \frac{1}{2}\boldsymbol{c}^\top\widehat{H}\boldsymbol{c} + \boldsymbol{c}^\top\widehat{\boldsymbol{g}}, \tag{4.41}$$

because of the relations $\widehat{f}(\boldsymbol{c}) = f(\boldsymbol{s})$ and $\|\boldsymbol{s}\|_M = \|\boldsymbol{c}\|_2$. Let $\boldsymbol{c}_{\rm opt}$ and $\boldsymbol{s}_{\rm opt}$ be the optimal solutions to (4.41) and (1.2), respectively. It can be seen that their associated Lagrangian multipliers $\lambda_{\rm opt}$ are the same. Notice that

$$\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g}) = M^{-1/2}\mathcal{K}_k(\widehat{H}, \widehat{\boldsymbol{g}}),$$

and therefore, with $\boldsymbol{s} = M^{-1/2}\boldsymbol{c}$, it holds that

$$\boldsymbol{s} \in \mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g}) \text{ with } \|\boldsymbol{s}\|_M \le \Delta \quad \Leftrightarrow \quad \boldsymbol{c} \in \mathcal{K}_k(\widehat{H}, \widehat{\boldsymbol{g}}) \text{ with } \|\boldsymbol{c}\|_2 \le \Delta.$$

Recalling that the iterate $\boldsymbol{s}_k$ from TLTRS for (1.2) is defined by (3.4), we have $\boldsymbol{s}_k = M^{-1/2}\boldsymbol{c}_k$, where

$$\boldsymbol{c}_k = \arg \min_{\boldsymbol{c} \in \mathcal{K}_k(\widehat{H}, \widehat{\boldsymbol{g}}),\ \|\boldsymbol{c}\|_2 \le \Delta} \widehat{f}(\boldsymbol{c}).$$

In the other word, $\boldsymbol{c}_k$ is the iterate from TLTRS applying to (4.41) for which our main results in Theorem 4.2 are applicable.

Theorem 4.4. *Let $\widehat{\theta}_1 \le \widehat{\theta}_2 \le \cdots \le \widehat{\theta}_n$ be the ordered eigenvalues of $\widehat{H} = M^{-1/2}HM^{-1/2}$, and $\lambda_{\rm opt}$ be the optimal Lagrange multiplier of the weighted-norm* TRS (1.2) *with a symmetric and positive definite matrix $M \in \mathbb{R}^{n\times n}$. Let the sequence $\{\boldsymbol{s}_k\}_{k=0}^{k_{\max}}$ be generated by* TLTRS *(with the preconditioned Lanczos process) for* (1.2)*.*

(i) *The sequence $\{f(\boldsymbol{s}_k)\}_{k=0}^{k_{\max}}$ is nonincreasing, and $f(\boldsymbol{s}_{k_{\max}}) = f(\boldsymbol{s}_{\rm opt})$ for the nondegenerate case, and*

$$f(\boldsymbol{s}_{k_{\max}}) + \frac{\widehat{\tau}^2\widehat{\theta}_1}{2} \le f(\boldsymbol{s}_{\rm opt}) \le f(\boldsymbol{s}_{k_{\max}})$$

for the degenerate case, where $\widehat{\tau}^2 = \Delta^2 - \|(H - \widehat{\theta}_1 M)^{\dagger}\boldsymbol{g}\|_M^2 \ge 0$ and $\widehat{\theta}_1 \le 0$;

(ii) *For the nondegenerate case, if $\|\boldsymbol{s}_{\rm opt}\|_M = \|\boldsymbol{s}_k\|_M = \Delta$ for some $0 \le k \le k_{\max}$, then*

$$0 \le f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt}) \le 2\|\widehat{H}_{\rm opt}\|_2\widehat{\zeta}_k^2, \tag{4.42a}$$

$$\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_M \le 2\sqrt{\widehat{\varkappa}}\,\widehat{\zeta}_k, \tag{4.42b}$$

where $\widehat{H}_{\rm opt} = \widehat{H} + \lambda_{\rm opt}I_n$, $\widehat{\boldsymbol{g}} = M^{-1/2}\boldsymbol{g}$, and

$$\widehat{\zeta}_k = \min\left\{\frac{\Delta}{\mathscr{T}_{k+1}(\widehat{\eta})}, \frac{2\|\widehat{\boldsymbol{g}}\|_2\epsilon_k^{\rm ra}(\widehat{\eta})}{\widehat{\theta}_n - \widehat{\theta}_1}\right\},$$

$$\widehat{\eta} = \frac{\widehat{\varkappa}+1}{\widehat{\varkappa}-1} \quad \textit{with} \quad \widehat{\varkappa} = \kappa(\widehat{H}_{\rm opt}) = \frac{\widehat{\theta}_n + \lambda_{\rm opt}}{\widehat{\theta}_1 + \lambda_{\rm opt}}.$$

$\epsilon_k^{\rm ra}(\widehat{\eta})$ is as defined in (4.23)*, and $\mathscr{T}_{k+1}(\widehat{\eta})$ is the $(k+1)$st Chebyshev polynomial of the first kind evaluated at $\widehat{\eta}$.*

Finally, it is worth mentioning that in order to obtain similar estimates for the bounds in (4.42) to the bounds in (4.35) for $M = I_n$, there are no additional costs needed except for computing the extreme Ritz values, i.e., the extreme eigenvalues of the tridiagonal matrix T_k resulting from the preconditioned Lanczos process [9, Algorithm 4.2] (or the preconditioned CG [9, Algorithm 4.1]). In fact, the preconditioned Lanczos process generates an M-orthonormal basis Q_k for $\mathcal{K}_k(M^{-1}H, M^{-1}\boldsymbol{g})$ and a tridiagonal T_k, satisfying (3.2). The Ritz values (i.e., the eigenvalues of T_k) serve as approximate eigenvalues for $\widehat{H}$, and the counterpart of (4.35) can be established. We omit the detail.

5 Numerical tests: sharpness of the upper bounds

In this first set of our numerical tests, we will test the sharpness of our established upper bounds in Theorem 4.2 and their estimates in Remark 4.3. This set consists of several numerical examples carried out in MATLAB (R2011b), and we choose medium size n and employ the built-in MATLAB routine, namely `trust`[5], for the original (1.1) as well as for the projected one (3.3). In fact, we will use $\boldsymbol{s}_{\text{opt}}$ by `trust` as the "exact" solution to compare against. Moreover, in order to control the numerical effects arising from the loss of orthogonality in Q_k from the Lanczos process, we use the *Lanczos algorithm with full reorthogonalization* [5, Algorithm 7.2] to generate Q_k. We shall compare the errors

$$f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}}) \quad \text{and} \quad \|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$$

with their corresponding bounds given in Theorem 4.2 and Remark 4.3.

In constructing the testing matrices H, without loss of generality we simply take H to be diagonal with translated Chebyshev nodes (to be defined) on the diagonal. We note that the n zero nodes and the $n+1$ extreme nodes of the nth Chebyshev polynomial $\mathscr{T}_n$ in $[-1,1]$ are given, respectively, by

$$t_{jn} = \cos\frac{(2j-1)\pi}{2n} \quad (1 \le j \le n) \quad \text{and} \quad \tau_{jn} = \cos\frac{j\pi}{n} \quad (0 \le j \le n).$$

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

$$z = \varpi(t - \xi) \tag{5.1}$$

maps $t \in [-1,1]$ one-to-one and onto $z \in [a,b]$, where

$$\varpi = \frac{b-a}{2} \quad \text{and} \quad \xi = -\frac{a+b}{b-a}.$$

The nth translated Chebyshev zero and extreme nodes on $[a,b]$ are defined to be the images of the nodes on $[-1,1]$ under the transformation (5.1), namely,

$$t^{\mathsf{tr}}_{jn} = \varpi(t_{jn} - \xi) \quad (1 \le j \le n) \quad \text{and} \quad \tau^{\mathsf{tr}}_{jn} = \varpi(\tau_{jn} - \xi) \quad (0 \le j \le n).$$

The reason behind choosing such matrices H is that the resulting linear systems $H_{\text{opt}}\,\boldsymbol{s} = -\boldsymbol{g}$ are the hardest for CG, MINRES, and GMRES for a fixed $\varkappa = \kappa(H_{\text{opt}})$ as confirmed by the theoretical analysis in [14, 15]. In summary, we take

$$H = \operatorname{diag}(t^{\mathsf{tr}}_{1n}, \ldots, t^{\mathsf{tr}}_{nn}), \quad \text{or} \tag{5.2a}$$

$$H = \operatorname{diag}(\tau^{\mathsf{tr}}_{0\,n-1}, \ldots, \tau^{\mathsf{tr}}_{n-1\,n-1}), \tag{5.2b}$$

$$\boldsymbol{g} = [1, \ldots, 1]^\top. \tag{5.2c}$$

In the TRS examples that follow, we will set various trust-region radii Δ and specify different intervals $[a,b]$ to construct the testing matrices H. They are indefinite, except the ones in Example 5.3. We will examine the observed $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$ together with their upper bounds by Theorem 4.2 and also the upper bound estimates given by (4.35).

[5] We point out that the built-in MATLAB routine `trust` is available in MATLAB 7.0 (R14). The solver `trust` is only suitable for small-to-medium size TRS as it solves the secular equation $1/\|\boldsymbol{s}(\lambda)\|_2 - 1/\Delta = 0$ by using the full eigen-decomposition of the coefficient matrix.

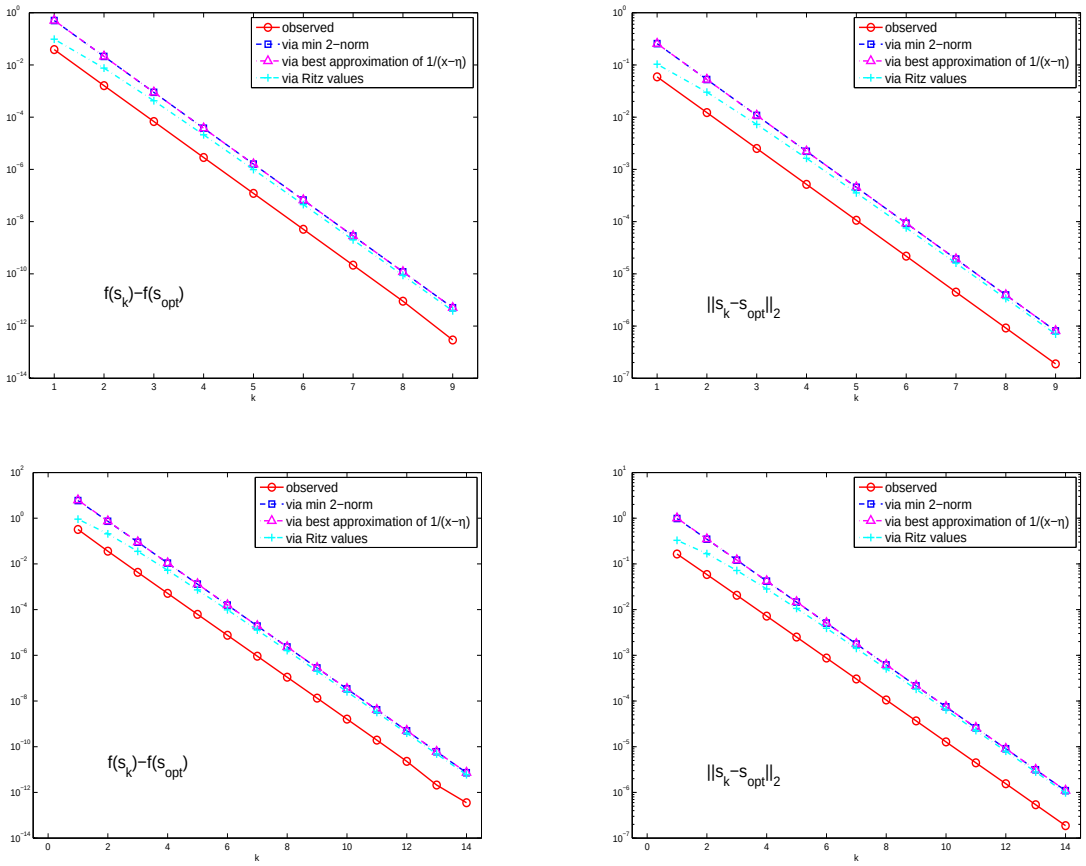

Figure 5.1: Example 5.1. *Top two plots*: $H \in \mathbb{R}^{500\times 500}$ with translated Chebyshev zero nodes on $[-10, 10]$; *bottom two plots*: $H \in \mathbb{R}^{500\times 500}$ with translated Chebyshev extreme nodes on $[-20, 20]$. The lines labelled by "via min 2-norm", "via best approximation of $1/(x-\eta)$", and "via Ritz values" are for (4.18), (4.24), and (4.35), respectively.

Example 5.1. Set $\Delta = 1$, $n = 500$. We test two different H: (5.2a) with $[a, b] = [-10, 10]$ and (5.2b) with $[a, b] = [-20, 20]$. By MATLAB's `trust`, we find

$[a, b]$	$\varkappa$	$\Gamma_\varkappa$	η	λ_{opt}	$f(\boldsymbol{s}_{\text{opt}})$	$\|\boldsymbol{s}_{\text{opt}}\|_2$
$[-10, 10]$	2.3006	4.8702	2.5378	25.3775	-23.4072	1.0000
$[-20, 20]$	4.2746	2.8749	1.6114	32.2276	-26.0120	1.0000

Additionally, we observed that the trust-region constraint is active when $k \geq 1$, i.e., $\dim \mathcal{K}_k(H, \boldsymbol{g}) \geq 2$. Figure 5.1 plots the observed $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$ together with the bounds according to (4.18) labelled as "*via min 2-norm*", (4.24) labelled as "*via best approximation of* $1/(x-\eta)$", and also the estimates given by (4.35) labelled as "*via Ritz values*". We observe from Figure 5.1 that

(1) the bounds for $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\text{opt}})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\text{opt}}\|_2$ given by (4.18) and (4.24) provide almost the same and quite satisfactory estimates, and there is no difference in convergence behavior for H by (5.2a) or by (5.2b), which is consistent with what is known for extreme examples for CG [14].

(2) interestingly, the estimates given by (4.35) provide sharper bounds for $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ than those by the (computed) eigenvalues of H. Also, with the Ritz values, the bounds (4.24) resulting from best polynomials for approximating $\frac{1}{x-\eta}$ are slightly sharper for $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ than that from (4.18). Overall, it shows that the estimates in Theorem 4.2 and Remark 4.3 are quite effective.

Example 5.2. In this example, we continue the numerical evaluation of the bounds in Theorem 4.2, and the estimates in (4.26) but let Δ vary. We use the 500th Chebyshev zero nodes (i.e., $[a,b]=[-1,1]$) to give H as in (5.2a) and $\boldsymbol{g}$ as in (5.2c). Again by MATLAB's `trust`, we find for $\Delta=50$ and 100

Δ	$\varkappa$	$\Gamma_\varkappa$	η	$\lambda_{\rm opt}$	$f(\boldsymbol{s}_{\rm opt})$	$\|\boldsymbol{s}_{\rm opt}\|_2$
50	12.4224	1.7922	1.1751	1.1751	-1.8740×10^3	50.0000
100	30.1620	1.4452	1.0686	1.0686	-6.0067×10^3	100.0000

In the left and right subfigures in Figure 5.2, we plot the observed $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ together with the upper bounds in Theorem 4.2 and also the estimates given by (4.35), with respect to k for the cases $\Delta=50$ and $\Delta=100$. The trust-region constraint is active when $k\geq1$, i.e., $\dim\mathcal{K}_k(H,\boldsymbol{g})\geq2$. In this example, (4.18) via the polynomials

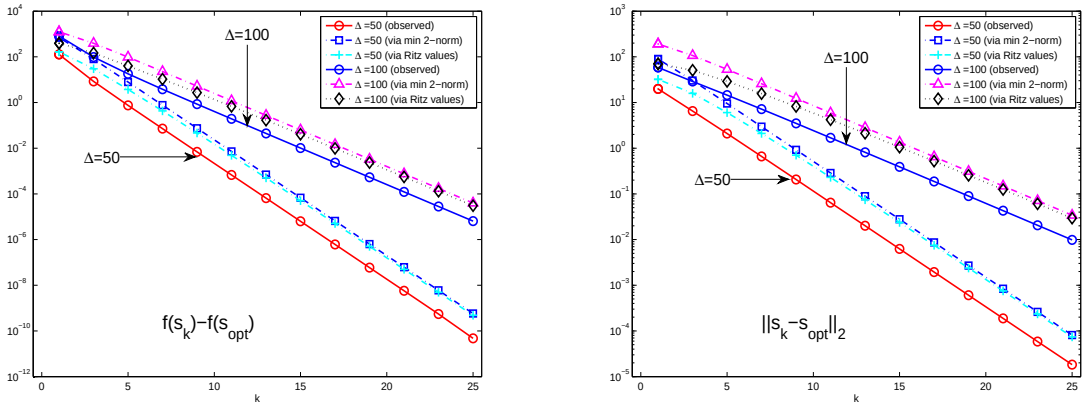

Figure 5.2: Example 5.2. $H\in\mathbb{R}^{500\times500}$ with the Chebyshev zero nodes on $[-1,1]$ and two trust-region radii $\Delta=50$ and $\Delta=100$. The lines labelled by "via min 2-norm" and "via Ritz values" are for (4.18) and (4.35), respectively.

resulting from $\min\|\boldsymbol{s}_{\rm opt}-\widetilde{\boldsymbol{s}}\|_2$ deliver the smaller bounds for $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ $(k=1,3,\ldots,25)$ of the two types of sub-optimal polynomials.

Also, we observed that $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ and their upper bounds go to zero faster as Δ gets smaller. This can be explained by Remark 4.2 and (4.33). In fact, the lower bound of η given by (4.31) implies that η decreases as Δ increases, which means that the smaller the radius Δ is, the faster the convergence.

Example 5.3. Lastly, we consider the situation when the interval $[a,b]$ varies. Following Example 5.2 and fixing $a=0$, $\Delta=1$ and $n=500$, we plot in Figure 5.3 the observed $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_2$ together with their upper bounds according to Theorem 4.2 and also the estimates given by (4.35), with respect to k for the two cases $b=50$ and

100. Again, the trust-region constraint becomes active when $k \geq 1$, i.e., $\dim \mathcal{K}_k(H, \boldsymbol{g}) \geq 2$. Relevant quantities for $b = 50$ and 100 are

b	$\varkappa$	$\Gamma_\varkappa$	η	$\lambda_{\rm opt}$	$f(\boldsymbol{s}_{\rm opt})$	$\|\boldsymbol{s}_{\rm opt}\|_2$
50	5.4382	2.5015	1.4506	11.2657	−15.1488	1.0000
100	12.4224	1.7922	1.1751	8.7544	−12.4794	1.0000

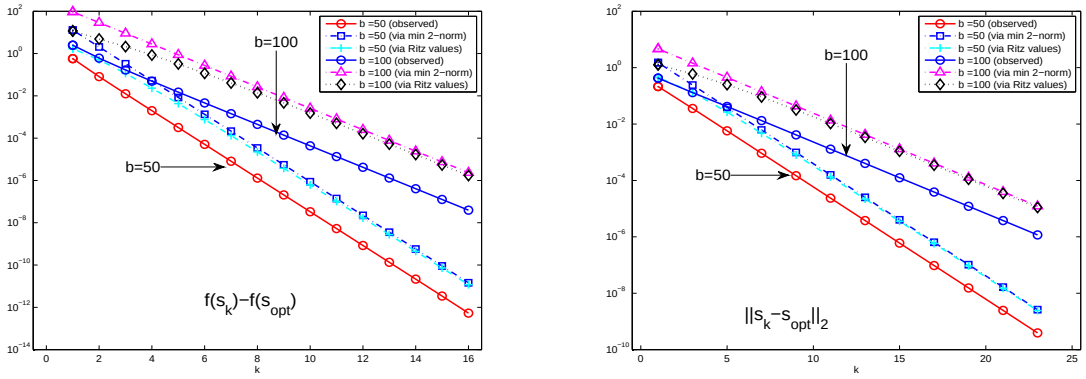

Figure 5.3: Example 5.3. $H \in \mathbb{R}^{500\times 500}$ with translated Chebyshev zero nodes on two intervals $[a, b] = [0, 50]$ and $[a, b] = [0, 100]$. The lines labelled by "via min 2-norm" and "via Ritz values" are for (4.18) and (4.35), respectively.

We point out that in this example, for both $b = 50$ and 100, we have

$$\zeta_k = \min\left\{\frac{\Delta}{\mathscr{T}_{k+1}(\eta)}, \frac{2\|\boldsymbol{g}\|_2\epsilon_k^{\rm ra}(\eta)}{\theta_n - \theta_1}\right\} = \frac{\Delta}{\mathscr{T}_{k+1}(\eta)}$$

for each computed Lanczos step k; in the other word, (4.18) via the polynomials resulting from $\min \|\boldsymbol{s}_{\rm opt} - \widetilde{\boldsymbol{s}}\|_2$ deliver the smaller values for $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2$ of the two types of sub-optimal polynomials.

Also, a similar pattern as in Figure 5.2 is observed: $f(\boldsymbol{s}_k) - f(\boldsymbol{s}_{\rm opt})$ and $\|\boldsymbol{s}_k - \boldsymbol{s}_{\rm opt}\|_2$ and their upper bounds go to zero faster as b get smaller. To explain it using Remark 4.2 and (4.33), we first note that $\lambda_{\rm opt}$ changes slightly for $b = 50$ and 100; moreover, $\varpi = \frac{b}{2}$ and

$$\theta_1 = \frac{b}{2}\left(1 + \cos\frac{(2n-1)\pi}{2n}\right) \approx \frac{b\pi^2}{16n^2}, \quad \theta_n = \frac{b}{2}\left(1 + \cos\frac{\pi}{2n}\right) \approx b - \frac{b\pi^2}{16n^2}.$$

Thus, $\eta \approx 1 + 2\frac{\lambda_{\rm opt}}{b}$ and since $\lambda_{\rm opt}$ changes slightly in the table above, η decreases when b increases from 50 to 100.

6 New stopping criteria

Gould, Orban and Toint [10] developed `GALAHAD` (version 2.6), a Fortran 2003 package for large scale nonlinear programming (available at `www.galahad.rl.ac.uk`). The solver `TRU` in `GALAHAD` uses the trust-region method (see [9, Algorithm 6.1]) to find a (local) unconstrained minimizer of a differentiable objective function. It offers direct and iterative

solvers for the related trust-region subproblem which is solved by the subroutine GLTR. TRU is most suitable for large scale problems.

GLTR uses the stopping criterion [9, Theorem 5.1]

$$\|(H+\lambda_k M)\boldsymbol{s}_k+\boldsymbol{g}\|_{M^{-1}}=\gamma_{k+1}|\boldsymbol{e}_{k+1}^{\top}\boldsymbol{h}_k|\le \mathtt{tol}_s, \tag{6.1}$$

where $\mathtt{tol}_s$ is a given tolerance. We propose to combine it with the upper bound for $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ given in (4.35); that is, we terminate the iteration in the second pass of GLTR if

$$\gamma_{k+1}|\boldsymbol{e}_{k+1}^{\top}\boldsymbol{h}_k|\le \mathtt{tol}_s \quad \text{or} \quad 2(\sigma_{k+1}^{(k)}+\lambda_k)\chi_k^2\le \mathtt{tol}_f\cdot(|f(\boldsymbol{s}_k)|+1) \tag{6.2}$$

is satisfied, where $\mathtt{tol}_f$ is another given tolerance, λ_k is the corresponding Lagrangian multiplier, and $\sigma_{k+1}^{(k)}$ and χ_k are the approximations to θ_n and ζ_k, respectively, as discussed in Remark 4.3. The upper bound for $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_{M^{-1}}$ given in (4.35) is not used here because on the one hand, this upper bound itself is derived from the bound for $f(\boldsymbol{s}_k)-f(\boldsymbol{s}_{\rm opt})$ (cf. (4.15)), and on the other hand, monitoring the accuracy of the objective function is the most critical issue in the framework of the trust-region method TRU (cf. [9, (6.2)]). In a way, the first inequality in (6.2) already takes $\|\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt}\|_{M^{-1}}$ into consideration because

$$\|(H+\lambda_k M)\boldsymbol{s}_k+\boldsymbol{g}\|_{M^{-1}}=\|(H+\lambda_{\rm opt}M)(\boldsymbol{s}_k-\boldsymbol{s}_{\rm opt})+(\lambda_k-\lambda_{\rm opt})M\boldsymbol{s}_k\|_{M^{-1}}.$$

Note $\gamma_{k+1}|\boldsymbol{e}_{k+1}^{\top}\boldsymbol{h}_k|\le \mathtt{tol}_s$ can numerically detect the breakdown in the Lanczos process. The second inequality in (6.2) controls how close $f(\boldsymbol{s}_k)$ is to the optimal values of the objective function in the worst-case scenario. Similarly to the classical convergence for the Krylov subspace method for the linear system (see e.g., [5, 28]), this upper bound cannot tell the occurrence of breakdown. This is another reason we choose to combine these two types of error bounds together as the new stopping criteria in GLTR.

Note from Remark 4.3 that only the smallest and largest eigenvalues of the tridiagonal matrix T_k are required to obtain $\sigma_{k+1}^{(k)}$ and χ_k used in (6.2). In our implementation, when $k\le 200$, we call the LAPACK [1] subroutine DSTERF to compute all eigenvalues of $T_k+\lambda_k I_{k+1}$, but for $k>200$, we call the subroutine GLTR_leftmost_eigenvalue built in gltr.f90 to compute the smallest eigenvalue $\sigma_1^{(k)}+\lambda_k$ of $T_k+\lambda_k I_{k+1}$, and employ a combination of the power iteration (as the first phase) and Rayleigh quotient iteration (as the second phase) to compute the largest eigenvalue $\sigma_{k+1}^{(k)}+\lambda_k$. In particular, suppose $(\mu_j,\boldsymbol{z}_j)$ is the approximate eigenpair after j power iterations. We are satisfied when the residual $\|T_k\boldsymbol{z}_j+\lambda_k\boldsymbol{z}_j-\mu_j\boldsymbol{z}_j\|_2\le 10^{-6}$; otherwise, as long as $j>100$, we then call at most 10 Rayleigh quotient iterations to refine the pair $(\mu_j,\boldsymbol{z}_j)$. To save the computational complexity in solving the smallest and largest eigenvalues of $T_k+\lambda_k I_{k+1}$, we choose to test the stopping criteria (6.2) only for odd k, i.e., every two iterations. Note that the reduced TRS (3.3) is solved for each k as in the original GLTR.

7 Numerical tests: effectiveness of new stopping criteria

GALAHAD [10] consists of a number of solvers for unconstrained and bound-constrained optimization, quadratic programming, nonlinear programming, systems of nonlinear equations and inequalities, and nonlinear least squares problems. We installed and compiled (under

gcc) GALAHAD and conducted our numerical testing on an iMac with 2.7GHz Intel Core i5, 8GB memory and OS X 10.9.5 system (64bit). The double precision arithmetic is used.

To test the trust-region method in TRU with the new stopping criteria (6.2) for GLTR, we integrate our upper bounds in Theorem 4.2 and Remark 4.3 into the subroutine gltr.f90 in GLTR.

Our test problems in this subsection are from the problem collection CUTEr[6]. Specifically, by specifying detailed options in the CUTEr collection, we systematically pick up and test all the unconstrained minimization problems from CUTEr. This yields a set of 86 test problems and the detailed options for extracting them are listed in the following table:

Objective function type	: Q S O
Constraints type	: U
Regularity	: *
Degree of available derivatives	: 2
Problem interest	: *
Explicit internal variables	: *
Number of variables	: [100, 99999999]
Number of constraints	: *

where Q = *quadratic type*, S = *sum of square type*, O = *other type (nonlinear, non-constant, etc.)*, U =*unconstrained*, * = *everything goes*, "Degree of available derivatives =2" means the explicit second-order Hessian is used, and "Number of variables =[100, 99999999]" means that the number of variables ranges from 100 to 99999999.

Important testing parameters are as follows.

(1) For the preconditioned matrix M, we noticed from the numerical testing in [9] that the unpreconditioned trust-region method (i.e., with $M = I_n$) often performs best overall. Therefore, in our numerical testing, we set $M = I_n$.

(2) The default options for the stopping criteria used in the package GLTR are $\texttt{tol}_s = 10^{-8}$, itmax $= n$ and the maximum CPU time maxcpu $= 1800$ secs. We remark that itmax is the maximum of the dimension k and the option itmax $= n$ implies that the Lanczos process is allowed to expand to the entire space $\mathbb{R}^n$ which could lead to severe loss of orthogonality in Q_k, and also increase remarkably computational burden in solving the reduced TRS problem (3.3). From this point of view, for GLTR, we choose and test two values itmax $= n$ and itmax $= \min\{n, 500\}$; also we increase the default maxcpu $= 1800$ secs to maxcpu $= 3600$ secs. For the modified GLTR with our new stopping criteria, we have evaluated the performance with various choices of $\texttt{tol}_s, \texttt{tol}_f$, itmax and maxcpu. The numerical results of TRU are presented in Tables A.1 and A.2 in the appendix, where we compare the original version of GLTR with the modified GLTR with $\texttt{tol}_s = 10^{-8}$, $\texttt{tol}_f = 0.005$ (Note itmax $= \min\{n, 500\}$ or itmax $= n$ and maxcpu $= 3600$ secs for both the original and modified GLTR).

(3) The labels '#it', '#g' and '#cg' in Tables A.1 and A.2 represent the total number of iterations in the trust-region method (TRU), the number of gradient evaluations for the objective function, and the total number of CG iterations required (which is equal

[6] www.cuter.rl.ac.uk/.

to the total Lanczos steps used in solving all involved trust-region subproblems), respectively. We remark that the total number of iterations #it is identical to the number of evaluations of the objective function.

According to our numerical experiences and the results in Tables A.1 and A.2, we make the following remarks:

(a) There are totally 76 problems in these tables; the other 10 problems out of the 86 problems are not listed because no method solves them successfully (in the sense that the default tolerance used in `TRU` is fulfilled), due to either reaching the maximum number of iterations (default is 10000) in `TRU` or reaching the maximum CPU time 3600 secs.

(b) With the new stopping criteria (6.2), `GLTR` (`itmax` $= \min\{n, 500\}$) solves all 76 problems successively, while `GLTR` (`itmax` $= n$) fails for the problem `EIGENBLS`; by contrast, the original `GLTR` fails for `EIGENBLS` and `EIGENCLS`.

(c) By excluding the problems `EIGENBLS` and `EIGENCLS`, we observed that in terms of CPU time, out of the 74 problems, the modified `GLTR` with the new stopping criteria (6.2) outperformed (in the sense of saving more than 5% CPU time) the original `GLTR` on 38 problems for `itmax` $= n$, and 34 problems for `itmax` $= \min\{n, 500\}$, respectively, while on only 6 problems for both `itmax` $= n$ and `itmax` $= \min\{n, 500\}$ the original `GLTR` won. Also, the CPU times for computing the Ritz values in the modified `GLTR` are relatively very small and negligible for most cases.

(d) For the 74 problems, the total CPU time savings from using the new stopping criteria (6.2) are 1044 secs for `itmax` $= n$ and 552 secs for `itmax` $= \min\{n, 500\}$. Though `TRU` with (6.2) totally requires 2944 (resp., 2167) more outer-loop iterations and 2406 (resp., 1901) more gradient evaluations for `itmax` $= n$ (resp., `itmax` $= \min\{n, 500\}$) than the default, the remarkable declines in CPU time primarily come from the reduction in the number of CG steps: totally, `TRU` with the new stopping criteria saves 873787 and 545744 CG steps for the cases `itmax` $= n$ and `itmax` $= \min\{n, 500\}$, respectively.

(e) It seems that `itmax` $= \min\{n, 500\}$ generally can give a better performance than `itmax` $= n$, but exceptions can happen, for example, in the problems `CURLY10`, `CURLY20`, `CURLY30`, `DIXON3DQ`, and `SPARSINE`.

To have a more detailed comparison on the results in Tables A.1 and A.2, we provide the performance profiles proposed by Dolan and Moré [6]. Let ς denote either the default `GLTR` or the one with the new stopping criteria (6.2), and $\mathcal{P}$ be the set consisting of 74 problems solved successfully. In terms of #it, for a particular solver ς and a test problem $\varrho \in \mathcal{P}$, we can compute

$$\upsilon = \log_2 \left(\frac{\#\text{it}(\varsigma, \varrho)}{\text{best } \#\text{it}(\varrho)} \right),$$

where "$\#\text{it}(\varsigma, \varrho)$" stands for the number of iterations that the solver ς takes on the problem ϱ and "best $\#\text{it}(\varrho)$" means the smallest number of iterations between the two solvers. With the value υ, we can roughly claim that, for the test problem ϱ, the solver ς is at worse 2^{υ}

times slower than the best in terms of the number of iterations. In Figure 7.1, we plot the curve

$$y_\varsigma(x) = \frac{1}{74} \times \text{size}\left\{\varrho \in \mathcal{P} : \log_2\left(\frac{\#\text{it}(\varsigma,\varrho)}{\text{best } \#\text{it}(\varrho)}\right) \leq x\right\}$$

with respect to x for both solvers ς. This provides a way to graphically compare the numbers of iterations from the default `GLTR` and the modified one. Similarly, we plot the performance profiles for #g, #cg, CPU time and ‖grad‖ in Figures 7.2, 7.3 , 7.4 and 7.5, respectively, where ‖grad‖ represents the norm of the gradient at the terminated point. Overall, these performance profiles indicate that, with the new stopping criteria in `GLTR`, the performance of `TRU` can be improved considerably on $\mathcal{P}$.

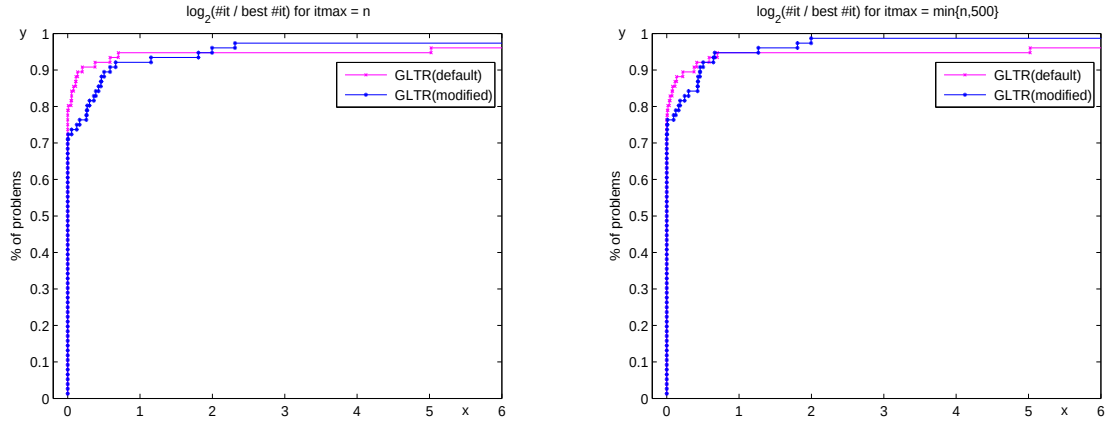

Figure 7.1: Performance profile for #it

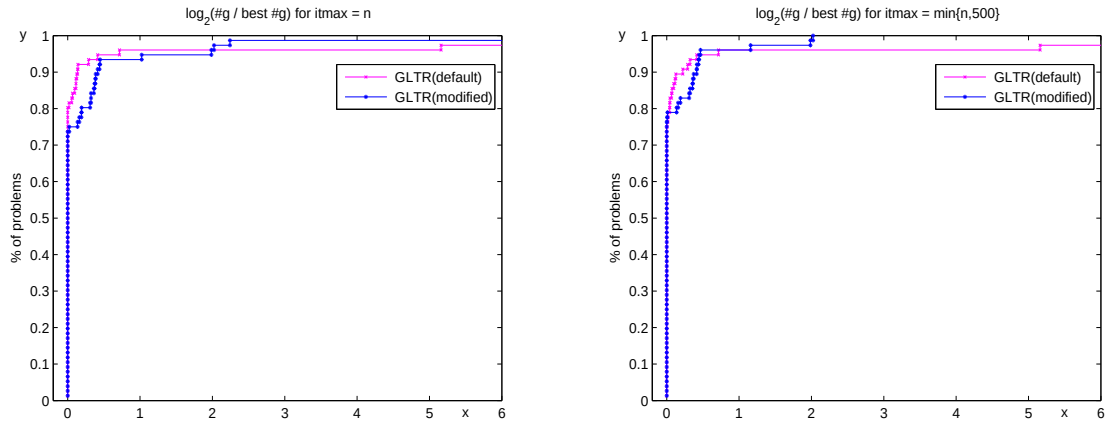

Figure 7.2: Performance profile for #g

As our final remark in this section, we explain why we chose a low accuracy $\texttt{tol}_f = 0.005$. We point out first that our newly-designed stopping criteria in the second inequality in (6.2) is to improve the overall performance of the trust-region method `TRU`. As observed and claimed in [9] that '*a more accurate approximation does not appear to significantly reduce the number of function evaluations within a standard trust-region method*', a suitable

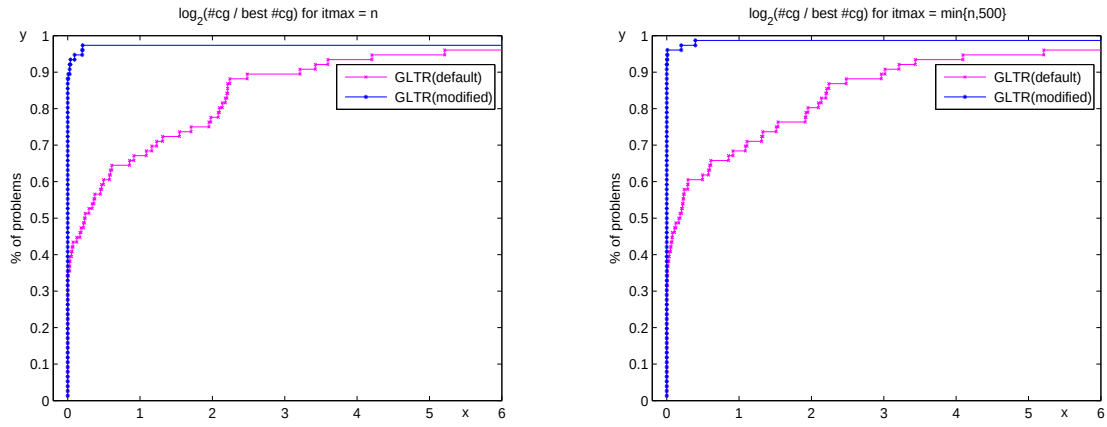

Figure 7.3: Performance profile for #cg

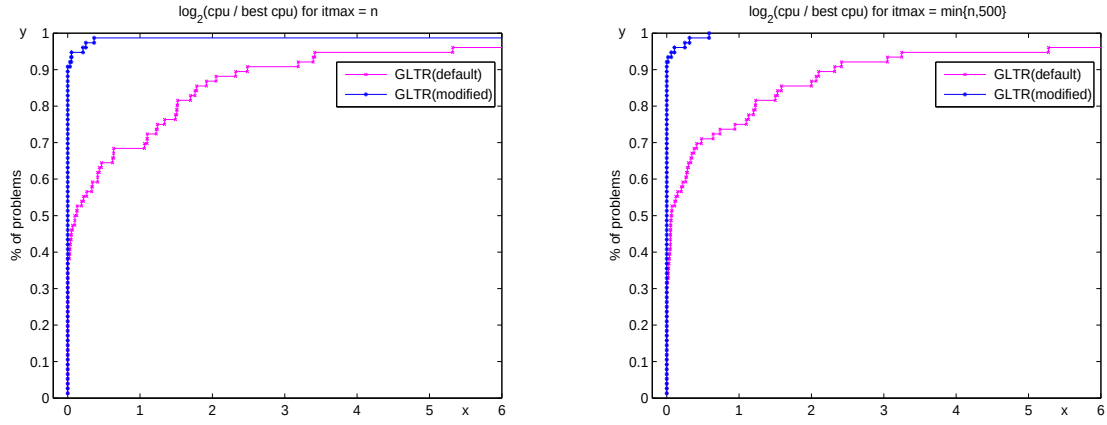

Figure 7.4: Performance profile for CPU time

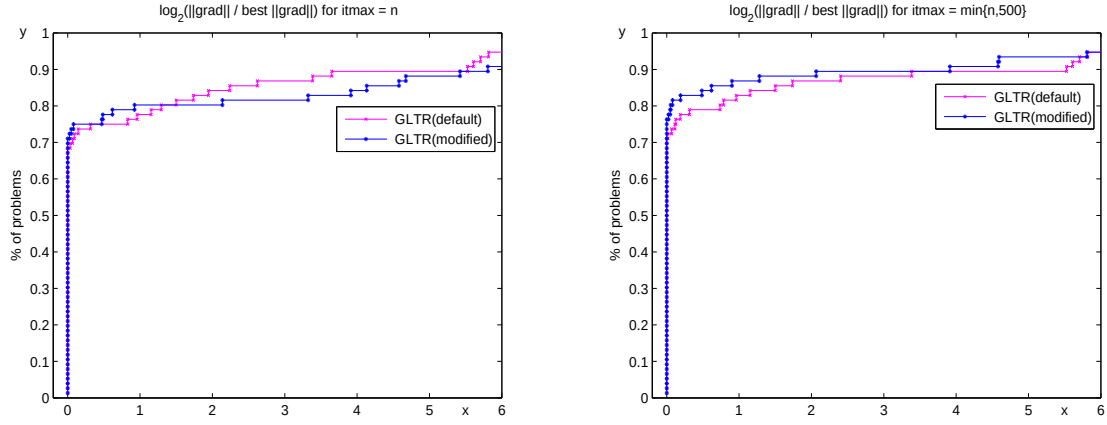

Figure 7.5: Performance profile for the norm of gradient $\|\text{grad}\|$.

low accuracy approximation (in terms of the objective function) is probably more appropriate for the overall performance of `TRU`. Indeed, our numerical experiments confirm the statement in [9], and a much smaller $\texttt{tol}_f$ could lead to the increase of the CPU time and

#cg for most problems. One may then argue that the accuracy $\gamma_{k+1}|\boldsymbol{e}_{k+1}^{\top}\boldsymbol{h}_k| \leq \texttt{tol}_s$ in the original stopping criteria can also be lowered down for a better performance of TRU. For this point, we have tested TRU with $\texttt{tol}_s = 10^{-6}$ and 10^{-4}, and the numerical results show that the overall performance of TRU can still be improved considerably by using the new stopping criteria (6.2). From this point of view, our upper bound developed in this paper is practically useful: it offers an estimate for the optimal objective function value, in contrast to the first inequality in (6.2) for the related KKT system alone, and provides a way to improve the performance of the trust-region solver TRU.

8 Concluding remarks

In this paper, we have performed convergence analysis for the generalized Lanczos Trust-Region (GLTR) method [9] which is an efficient implementation of the truncated Lanczos approach (TLTRS). Mimicking the classical Rayleigh-Ritz procedure in the eigenvalue computations, TLTRS first projects a large-scale TRS (1.2) into a much smaller TRS (3.3) using the (preconditioned) Lanczos process, and then solves the smaller TRS (3.3) by the Moré-Sorensen algorithm or some modifications of it. It is interesting to point out that, when $M = I_n$ and $\boldsymbol{g} = 0$, TRS reduces to the standard symmetric eigenvalue problem which can be solved by the Lanczos method (e.g., [5, 22]). In that special case, the global optimal value $f(\boldsymbol{s}_{\rm opt})$ is the smallest eigenvalue θ_1 of H while the global optimal value $f(\boldsymbol{s}_k)$ of the projected problem (3.3) is the smallest Ritz value σ_1. Elegant theoretical *a priori* error bounds concerning the eigenvalues and eigenvectors approximated by the Ritz values and Ritz vectors by the Lanczos method have been established (e.g., [17, 22, 27]). The Chebyshev polynomials of the first kind have been playing a critical role in these elegant theoretical results.

By making use of the special structure and optimality conditions of TRS, this paper addresses the theoretical question: how good is the projected TRS in approximating the original problem? We have established *a priori* error bounds on the differences between the approximate objective value $f(\boldsymbol{s}_k)$ (cf. Ritz value) as well as the approximate solution (cf. Ritz vector) $\boldsymbol{s}_k$ and the corresponding optimal ones. It is interesting to point out that, besides the Chebyshev polynomials of the first kind, the best polynomial approximations of the rational function $\frac{1}{x-\eta}$ in the interval $[-1, 1]$ also play a role in characterizing the convergence behavior of TLTRS. Our error bounds turn out to be rather sharp in general and can be numerically estimated at roughly $O(k^2)$ flops. Most importantly, the estimates can be used to devise stopping criteria for TLTRS/GLTR, and indeed we have proposed new stopping criteria which has been integrated into GLTR in the library GALAHAD. Numerical examples are presented to support our theoretical analysis and test the trust-region solver TRU and a modification of it to include our new stopping criteria on problems from CUTEr. Although we don't have a rigorous mathematical proof, our numerical results show that, with the new stopping criteria, the overall performance of the trust-region solver TRU is considerably improved.

Acknowledgements

The authors are grateful to associate editor Prof. William Hager, Prof. Nick Gould, and anonymous referees for their careful reading, useful comments, and suggestions which significantly improved the presentation of the paper.

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 3rd edition, 1999.

[2] S. N. Bernstein. Sur l'ordre de la meilleure approximation des fonctions continues par les polynômes de degré donné. *Mém. acad. royale Belg.*, 4:1–104, 1912.

[3] E. W. Cheney. *Introduction to Approximation Theory*. Chelsea Publishing Company, New York, 2nd edition, 1982.

[4] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, PA, 2000.

[5] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

[6] E. D. Dolan and J. Moré. Benchmarking optimization software with performance profiles. *Math. Progr.*, 91:201–213, 2002.

[7] D. M. Gay. Computing optimal locally constrained steps. *SIAM J. Sci. Statist. Comput.*, 2(1):186–197, 1981.

[8] G. H. Golub and U. von Matt. Quadratically constrained least squares and quadratic problems. *Numer. Math.*, 59:561–580, 1991.

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

[10] 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(4):353–372, 2004.

[11] N. I. M. Gould, D. P. Robinson, and H. S. Thorne. On solving trust-region and other regularised subproblems in optimization. *Math. Progr. Comput.*, 2(1):21–57, 2010.

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

[13] W. W. Hager and Y. Krylyuk. Graph partitioning and continuous quadratic programming. *SIAM J. Alg. Discrete Methods*, 12:500–523, 1999.

[14] R.-C. Li. Vandermonde matrices with Chebyshev nodes. *Linear Algebra Appl.*, 428:1803–1832, 2007.

[15] R.-C. Li. On Meinardus' examples for the conjugate gradient method. *Math. Comp.*, 77(261):335–352, 2008. Electronically published on September 17, 2007.

[16] R.-C. Li. Sharpness in rates of convergence for symmetric Lanczos method. *Math. Comp.*, 79(269):419–435, 2010.

[17] R.-C. Li and L.-H. Zhang. Convergence of block Lanczos method for eigenvalue clusters. *Numer. Math.*, 131:83–113, 2015.

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

[19] G. Meinardus. *Approximation of Functions: Theory and Numerical Methods. Translated by L. L. Schumaker.* Berlin Heidelberg, Springer, 1967.

[20] J. Moré and D. Sorensen. Computing a trust region step. *SIAM J. Sci. Statist. Comput.*, 4(3):553–572, 1983.

[21] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[22] B. N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, 1998.

[23] R. Rendl and H. Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Math. Progr.*, 77(2):273–299, 1997.

[24] M. Rojas, S. A. Santos, and D. C. Sorensen. A new matrix-free algorithm for the large-scale trust-region subproblem. *SIAM J. Optim.*, 11:611–646, 2000.

[25] 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(2):11:1–28, 2008.

[26] M. Rojas and D. C. Sorensen. A trust-region approach to the regularization of large-scale discrete forms of ill-posed problems. *SIAM J. Sci. Comput.*, 23:1843–1861, 2002.

[27] Y. Saad. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM J. Numer. Anal.*, 15(5):687–706, October 1980.

[28] Y. Saad. *Iterative Methods for Linear Systems*. PWS, Boston, USA, 1996.

[29] D. C. Sorensen. Newton's method with a model trust region modification. *SIAM J. Numer. Anal.*, 19(2):409–426, 1982.

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

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

[32] A. Tarantola. *Inverse Problem Theory*. Elsevier, Amsterdam, The Netherlands, 1987.

[33] A. N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Math.*, 4:1624–1627, 1963.

[34] P. L. Toint. *Towards an efficient sparsity exploiting Newton method for minimization*. In Sparse Matrices and Their Uses, Academic Press, London, 57-88, I. Duff, edition, 1981.

[35] Y. Yuan. On the truncated conjugate gradient method. *Math. Progr.*, 87:561–573, 2000.

[36] L.-H. Zhang, W. H. Yang, C. Shen, and J. Feng. Error bounds of the Lanczos approach for the trust-region subproblem. Technical report, May 2015. DOI: 10.13140/RG.2.1.1846.2808, Available at `www.researchgate.net/publication/289254863`.

[37] Y. Zhou and R.-C. Li. Bounding the spectrum of large Hermitian matrices. *Linear Algebra Appl.*, 435:480–493, 2011.

A Numerical results of `GLTR` and the modified `GLTR`

Table A.1: A comparison of GLTR (default) with the modified GLTR using the new stopping criteria on CUTEr problems.

Prob.	n	itmax $= n$								itmax $= \min\{n, 500\}$							
		GLTR (default)				GLTR (new stopping criteria)				GLTR (default)				GLTR (new stopping criteria)			
		#it	#g	#cg	cpu	#it	#g	#cg	cpu	#it	#g	#cg	cpu	#it	#g	#cg	cpu
ARGLINA	200	2	3	2	0.30	2	3	2	0.28	2	3	2	0.29	2	3	2	0.28
ARGLINB	200	2	3	2	0.28	2	3	2	0.28	2	3	2	0.29	2	3	2	0.28
ARGLINC	200	2	3	2	0.27	2	3	2	0.28	2	3	2	0.28	2	3	2	0.27
ARWHEAD	5000	6	7	12	0.01	6	7	12	0.01	6	7	12	0.01	6	7	12	0.01
BDQRTIC	5000	9	10	169	0.06	9	10	169	0.05	9	10	169	0.06	9	10	169	0.05
BOX	10000	7	7	16	0.03	7	7	16	0.03	7	7	16	0.03	7	7	16	0.03
BROWNAL	200	4	5	16	0.30	4	5	16	0.28	4	5	16	0.29	4	5	16	0.28
BROYDN7D	5000	247	214	33078	3.90	277	244	7148	1.36	247	214	33078	3.96	277	244	7148	1.37
BRYBND	5000	16	14	883	0.21	22	18	1015	0.25	16	14	883	0.21	22	18	1015	0.25
CHAINWOO	4000	484	404	165469	13.11	2403	1912	190954	16.89	1779	1432	479202	32.86	2403	1912	190954	17.10
COSINE	10000	325	322	4665	1.60	10	9	126	0.04	325	322	4665	1.55	10	9	126	0.04
CRAGGLVY	5000	17	18	754	0.12	17	18	754	0.11	17	18	754	0.12	17	18	754	0.11
CURLY10	10000	23	22	152373	121.02	20	20	110074	56.53	2794	2793	1395535	770.12	2783	2783	1388504	738.32
CURLY20	10000	27	26	164526	192.56	26	24	127998	125.20	3097	3096	1546070	1536.93	3109	3107	1549764	1604.03
CURLY30	10000	27	26	163109	261.36	26	25	125737	168.32	3291	3290	1643423	2415.02	3313	3312	1651607	2350.47
DIXMAANA	3000	6	7	24	0.01	6	7	22	0.01	6	7	24	0.01	6	7	22	0.01
DIXMAANB	3000	11	11	777	0.05	11	11	84	0.01	11	11	777	0.05	11	11	84	0.01
DIXMAANC	3000	12	11	3041	0.21	17	15	165	0.02	12	11	2816	0.19	17	15	165	0.02
DIXMAAND	3000	12	12	3256	0.32	19	16	269	0.03	12	12	2102	0.16	19	16	269	0.03
DIXMAANE	3000	11	10	2240	0.14	12	11	902	0.06	11	10	2240	0.14	12	11	902	0.06
DIXMAANF	3000	30	24	3729	0.22	37	30	2498	0.21	30	24	3729	0.23	37	30	2498	0.20
DIXMAANG	3000	31	26	4764	0.29	30	24	854	0.07	31	26	4764	0.30	30	24	854	0.07
DIXMAANH	3000	29	25	5438	0.34	39	31	1150	0.09	29	25	5438	0.36	39	31	1150	0.09
DIXMAANI	3000	15	14	28856	4.86	18	16	30834	1.73	35	35	14784	0.98	30	30	12088	0.77
DIXMAANJ	3000	34	28	21499	1.51	47	35	4641	0.27	44	36	17490	1.13	47	35	4641	0.27
DIXMAANK	3000	36	30	28947	2.54	80	61	12927	0.75	42	34	18610	1.21	101	76	16577	0.96
DIXMAANL	3000	41	34	44441	5.19	49	38	10505	0.57	49	40	22189	1.49	49	38	2742	0.18
DIXON3DQ	10000	2	3	11417	1.77	2	3	10147	1.52	1054	1055	527499	92.43	1042	1043	520647	88.08
DQDRTIC	5000	3	4	17	0.01	2	3	8	0.00	3	4	17	0.01	2	3	8	0.00
DQRTIC	5000	29	30	8752	0.60	29	30	8924	0.58	29	30	8071	0.54	29	30	8018	0.52
EDENSCH	2000	16	16	275	0.02	16	16	242	0.02	16	16	275	0.02	16	16	242	0.02
EG2	1000	3	4	3	0.00	3	4	3	0.00	3	4	3	0.00	3	4	3	0.00
EIGENALS	2550	52	46	10388	169.13	48	43	7366	122.38	52	46	10388	169.84	48	43	7366	121.80
EIGENBLS	2550	-	-	-	-	-	-	-	-	-	-	-	-	649	502	174683	2644.72
EIGENCLS	2652	-	-	-	-	585	476	204427	3358.86	-	-	-	-	589	481	119529	1898.48
ENGVAL1	5000	9	10	107	0.02	9	10	91	0.02	9	10	107	0.02	9	10	91	0.02
EXTROSNB	1000	1050	797	18657	0.53	1044	797	17890	0.53	1050	797	18657	0.53	1044	797	17890	0.51
FLETCBV2	5000	0	1	0	0.00	0	1	0	0.00	0	1	0	0.00	0	1	0	0.00

Table A.2: A comparison of GLTR (default) with the modified GLTR using the new stopping criteria on CUTEr problems.

Prob.	n	itmax $= n$								itmax $= \min\{n, 500\}$							
		GLTR (default)				GLTR (new stopping criteria)				GLTR (default)				GLTR (new stopping criteria)			
		#it	#g	#cg	cpu	#it	#g	#cg	cpu	#it	#g	#cg	cpu	#it	#g	#cg	cpu
FLETCHCR	1000	1417	1415	70655	1.58	1417	1415	59747	1.54	1417	1415	70655	1.84	1417	1415	59747	1.52
FMINSRF2	5625	307	299	28451	5.62	1224	1214	6473	6.51	307	299	28451	5.71	1224	1214	6473	6.15
FMINSURF	5625	84	73	13056	1035.12	294	289	3052	666.73	84	73	13056	1038.07	294	289	3052	665.43
FREUROTH	5000	9	10	76	0.03	9	10	62	0.03	9	10	76	0.03	9	10	62	0.03
GENROSE	500	438	353	33798	0.38	414	337	18691	0.30	438	353	33798	0.39	414	337	18691	0.30
JIMACK	3549	28	23	106296	70.91	36	30	45336	28.02	46	44	26242	19.85	62	57	24850	20.09
LIARWHD	5000	13	14	29	0.02	13	14	29	0.02	13	14	29	0.02	13	14	29	0.02
MANCINO	100	11	12	271	0.16	10	11	59	0.14	11	12	271	0.16	10	11	59	0.13
MODBEALE	20000	14	14	337	0.34	14	14	325	0.34	14	14	337	0.34	14	14	325	0.33
MOREBV	5000	1	2	4434	0.47	1	2	4434	0.46	1	2	500	0.05	1	2	500	0.05
MSQRTALS	1024	41	32	48140	110.57	38	29	14761	34.08	41	33	25127	59.73	40	32	11647	27.28
MSQRTBLS	1024	34	26	38084	86.82	41	34	13048	30.56	32	25	20614	48.98	38	32	7098	17.30
NCB20	5010	76	66	24094	12.66	76	67	5317	3.68	80	69	15175	8.29	76	67	5317	3.61
NCB20B	5000	16	14	7884	3.78	24	19	5763	3.00	16	14	6567	3.21	25	19	5699	2.97
NONCVXU2	5000	2085	1760	227074	32.28	2094	1738	57634	13.67	2113	1788	221043	31.85	2121	1765	58229	13.58
NONCVXUN	5000	2014	1769	663977	96.17	2091	1770	525063	71.07	2331	2086	399753	55.96	2648	2327	370817	53.26
NONDIA	5000	4	5	10	0.01	4	5	10	0.01	4	5	10	0.01	4	5	10	0.01
NONDQUAR	5000	18	19	48085	5.08	18	19	47103	4.90	18	19	7041	0.78	18	19	6022	0.64
OSCIGRAD	100000	11	10	443	1.20	11	10	443	1.20	11	10	443	1.20	11	10	443	1.18
PENALTY1	1000	26	27	26	0.69	26	27	26	0.63	26	27	26	0.70	26	27	26	0.70
PENALTY2	200	10	11	753	0.07	10	11	753	0.07	10	11	753	0.07	10	11	753	0.06
PENALTY3	200	16	16	420	0.45	16	16	361	0.46	16	16	420	0.46	16	16	361	0.45
POWELLSG	5000	18	19	81	0.02	18	19	77	0.02	18	19	81	0.02	18	19	77	0.02
POWER	10000	18	19	6677	1492.93	18	19	6653	1462.61	18	19	6677	1475.97	18	19	6653	1450.47
QUARTC	5000	29	30	8752	0.57	29	30	8924	0.59	29	30	8071	0.53	29	30	8018	0.52
SCHMVETT	5000	4	5	180	0.04	4	5	118	0.03	4	5	180	0.04	4	5	118	0.03
SENSORS	100	26	23	1416	0.15	16	14	132	0.07	26	23	1428	0.15	16	14	132	0.07
SINQUAD	5000	9	10	38	0.02	9	10	38	0.02	9	10	38	0.02	9	10	38	0.03
SPARSINE	5000	100	85	455936	254.14	131	110	302662	121.71	795	785	403165	171.17	1095	1085	530311	212.78
SPARSQUR	10000	18	19	1102	1.18	18	19	1108	1.16	18	19	1102	1.14	18	19	1108	1.13
SPMSRTLS	4999	26	22	4317	0.57	20	18	1115	0.20	26	22	4317	0.57	20	18	1115	0.19
SROSENBR	5000	6	7	13	0.01	6	7	13	0.01	6	7	13	0.01	6	7	13	0.01
TESTQUAD	5000	3	4	1976	0.13	3	4	2029	0.13	4	5	1513	0.10	3	4	1003	0.06
TOINTGSS	5000	10	10	166	0.04	10	10	88	0.03	10	10	166	0.03	10	10	88	0.03
TQUARTIC	5000	10	10	37	0.01	10	10	37	0.01	10	10	37	0.01	10	10	37	0.01
TRIDIA	5000	3	4	1203	0.11	3	4	1178	0.11	3	4	1027	0.10	3	4	1003	0.09
VARDIM	200	18	19	18	0.01	18	19	18	0.01	18	19	18	0.01	18	19	18	0.01
WOODS	4000	48	43	249	0.05	48	43	244	0.05	48	43	249	0.05	48	43	244	0.05