On the Convergence of the Truncated Lanczos Approach for the Trust-Region Subproblem

Lei-Hong Zhang
Ren-Cang Li

Technical Report 2015-14

http://www.uta.edu/math/preprint/
On the Convergence of the Truncated Lanczos Approach for the Trust-Region Subproblem

Lei-Hong Zhang* Ren-Cang Li†

July 30, 2015

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 truncated Lanczos approach proposed in [Gould, Lucidi, Roma and Toint, SIAM J. Optim., 9:504–525 (1999)] is a natural extension of the classical Lanczos method for the symmetric linear system and the symmetric eigenvalue problem to the trust-region subproblem. It mimics the classical Rayleigh-Ritz procedure for eigenvalue computations and is suitable for large scale problems. In this paper, we analyze the convergence of the truncated Lanczos approach and reveals its convergence behavior in theory. In particular, we develop quite sharp upper bounds for both the optimal objective value as well as the optimal solution. These bounds can be numerically estimated in an economical way to serve as practical stopping criteria for the truncated Lanczos approach. Numerical examples are reported to support our analysis.

Key words. Trust-region method, Trust-region subproblem, Steihaug–Toint conjugate-gradient iteration, Lanczos method, Rayleigh-Ritz procedure

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

1 Introduction

Minimization of a quadratic function over a Euclidean ball

$$\min_{\|s\|_2 \leq \Delta} f(s) \quad \text{with} \quad f(s) := \frac{1}{2} s^\top H s + s^\top g$$  \hfill (1.1)
is widely known as the trust-region subproblem (TRS) [16, 17], where $H = H^T \in \mathbb{R}^{n \times n}$ and $g \in \mathbb{R}^n$. It also shows up in other important applications such as the Tikhonov regularization [20, 21, 22, 28, 29] for ill-posed problems, graph partitioning problems [9] as well as in the Levenberg-Marquardt approach in optimization [17].

Because of its vital role in numerous applications, several algorithms have been proposed to solve (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., [16, 17]) and factorization-free algorithms for large-scale sparse problems (see, e.g., [5, 6, 8, 17, 19, 20, 21, 22, 26, 27, 28, 30]).

The Moré-Sorensen method [16] available as subroutine GQTPAR in MINPACK-2 is probably the most well-known one for small size dense problems, and it or its modifications are frequently embedded into programs as a building block for solving relevant subproblems within large-scale computational problems. This is also the case for the method proposed in [6] (see also [3, Chapter 5]). In particular, the authors of [6] presented a generalized Lanczos trust-region method (GLTR) [6, Algorithm 5.1] which is an improved Steihaug–Toint [27, 30] conjugate gradient (CG) iteration for the weighted-norm trust-region subproblem:

$$\min_{\|s\|_M \leq \Delta} f(s),$$

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

GLTR starts with the preconditioned CG iteration [6, Algorithm 4.1] for minimizing $f(s)$ and can automatically detect whether the solution, namely $s_{\text{opt}}$, to (1.2) lies strictly inside the trust region $\|s\|_M \leq \Delta$ or on its boundary $\|s\|_M = \Delta$. In particular, as the CG iteration progresses, it can tell the boundary case $\|s_{\text{opt}}\|_M = \Delta$ either when the piecewise linear path connecting the CG iterates leaves the trust region $\|s\|_M \leq \Delta$, or a direction with negative curvature (a vector $p$ is a direction of negative curvature if $p^T Hp < 0$) is found. The latter implies $H$ is indefinite and $\|s_{\text{opt}}\|_M = \Delta$. By making use of the intimate relationship between the CG iteration and the Lanczos process (see [6, Section 4]), as soon as $\|s_{\text{opt}}\|_M = \Delta$ is detected, GLTR needs to solve smaller size trust-region subproblems afterwards, which are resulted from projecting the original TRS (1.2) onto the Krylov subspace generated by the Lanczos process, or equivalently, by CG (see [6] in detail). Extensive numerical testing suggests that this method 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.

The truncated Lanczos approach (abbreviated as TLTRS in this paper) [6, Section 5] is a slight modification of GLTR. TLTRS mimics the classical Rayleigh-Ritz procedure (see [18, Section 11.3] and [4, Definition 7.1]) for the eigenvalue problem and proceeds iteratively the following three steps: for $k = 0, 1, \ldots$ (the detailed procedure will be described in section 3):

1. generate the $k$th Krylov subspace by the preconditioned Lanczos process [18, Algorithm 4.2];

2. project the original TRS (1.2) onto the $k$th Krylov subspace to give a smaller size TRS;
3. solve the resulting smaller size TRS to get an approximate solution of TRS (1.2).

Both GLTR and TLTRS can be viewed as natural extensions of the classical Lanczos method for the linear system (when $g \neq 0$) and the eigenvalue problem (when $g = 0$) to TRS. There has been a wealth of developments, in both theory and implementation, on the Lanczos-based methods, e.g., in [4, 18, 24] for a complete development up to 1998 and more recently in [11, 12, 13]. However, to the authors’ best knowledge, convergence analysis for the Lanczos type method for TRS has not yet been fully developed. This paper attempts to analyze the convergence of TLTRS and GLTR and to understand the numerical behavior in theory. In particular, we shall develop sharp upper bounds for both the optimal objective value as well as the optimal solution; also, practically effective approximations to these upper bounds can be achieved using roughly $O(k^2)$ extra flops, which turns out to be very practical as $k \ll n$ in general, and can therefore be used to devise good stopping criteria for TLTRS and GLTR. Numerical examples are reported to support our claims.

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 describe the framework of TLTRS as well as some basic properties. Section 4 contains the main results of this paper: in subsections 4.1 and 4.2, we discuss the convergence of TLTRS for the case $\lambda_{\text{opt}} = 0$ and $\lambda_{\text{opt}} \neq 0$, respectively, where $\lambda_{\text{opt}}$ denotes the Lagrangian multiplier of (1.1) associated with the solution $s_{\text{opt}}$; subsection 4.3 shows how to extend the main convergence results to the weighted-norm TRS (1.2). Our numerical examples support our theoretical analysis are presented in section 5. Finally we conclude this paper in section 6 with a connection to error bounds for the eigenvalues/eigenvectors with respect to the Ritz values/vectors.

Notation. Throughout this paper, all vectors are column vectors and are typeset in bold lower case letters. For $x \in \mathbb{R}^n$ (the set of all real $n$-vectors), $x_i$ stands for its $i$th entry. For $A \in \mathbb{R}^{n \times m}$ (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 $e_j$ is the $j$th 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 $\ell$ and column $i$ to column $j$.

2 Optimality Conditions

The following well-known optimality conditions are due to Moré and Sorensen [16] (see also [25] and [17, Theorem 4.1]). It has been serving as the fundamental guideline for most existing methods for TRS.

Theorem 2.1 ([25]). The vector $s_{\text{opt}}$ is a global optimal solution of the trust-region problem (1.1) if and only if $s_{\text{opt}}$ is feasible, i.e., $\|s_{\text{opt}}\|_2 \leq \Delta$, and there is a scalar $\lambda_{\text{opt}} \geq 0$. 

such that the following conditions are satisfied:

\[
\begin{align*}
(H + \lambda_{\text{opt}} I_n) s_{\text{opt}} &= -g, & \lambda_{\text{opt}}(\Delta - \|s_{\text{opt}}\|_2) &= 0, & \text{and} \\
H + \lambda_{\text{opt}} I_n &\text{ is positive semidefinite.}
\end{align*}
\]

Let the eigen-decomposition of \(H\) be

\[
H = U \text{ diag}(\theta_1, \theta_2, \ldots, \theta_n) U^\top =: U \Theta U^\top,
\]

where the eigenvector matrix \(U \equiv [u_1, u_2, \ldots, u_n]\) is orthogonal, and

\[
\theta_1 = \theta_2 = \cdots = \theta_p < \theta_{p+1} \leq \cdots \leq \theta_n \tag{2.1}
\]

are the eigenvalues. In (2.1), we assume \(\theta_1\) has multiplicity \(p\). Let \(E_1\) be the invariant subspace associated with the smallest eigenvalue \(\theta_1\). Then \(U_1 := [u_1, \ldots, u_p] \in \mathbb{R}^{n \times p}\) is an orthonormal basis matrix for \(E_1\). Write \(U = [U_1, U_2]\), where \(U_2 = [u_{p+1}, \ldots, u_n]\), and set \(E_2 = \mathcal{R}(U_2) = E_1^\perp\), the orthogonal complement of \(E_1\).

For TRS (1.1), there are two situations (see, e.g., [8, 16, 17]) to consider:

1. the degenerate case [8, Lemma 2.2] (or the hard case [17]) as characterized by

\[
g \perp E_1 \text{ and } \|(H - \theta_1 I_n)^\dagger g\|_2 \leq \Delta \tag{2.2}
\]

and the corresponding Lagrangian multiplier is \(\lambda_{\text{opt}} = -\theta_1\). In this case, there are multiple global solutions which can be expressed by [8, Lemma 2.2]

\[
s_{\text{opt}} = -(H - \theta_1 I_n)^\dagger g + \tau u \tag{2.3}
\]

for any \(u \in E_1\) with \(\|u\|_2 = 1\), and

\[
\tau^2 = \Delta^2 - \|(H - \theta_1 I_n)^\dagger g\|_2^2 > 0;
\]

2. the nondegenerate case [8, Lemma 2.2] (or the easy case [17]) as characterized by the opposite of (2.2). In this case, the corresponding Lagrangian multiplier \(\lambda_{\text{opt}} > -\theta_1\), and the global solution \(s_{\text{opt}}\) is unique and given by [8, Lemma 2.2]

\[
s_{\text{opt}} = -(H + \lambda_{\text{opt}} I_n)^{-1} g.
\]

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

\[\text{footnote}{1}\text{We adopt the definitions of degenerate and nondegenerate cases used in [8, Lemma 2.2] in this paper.}\]
3 Truncated Lanczos for TRS (TLTRS)

We first outline TLTRS method [6, 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 matrix \( Q_k = [q_0, q_1, \ldots, q_k] \in \mathbb{R}^{n \times (k+1)} \) of the \((k+1)\)st Krylov subspace\(^2\)

\[
\mathcal{K}_k(M^{-1}H, M^{-1}g) := \mathbb{R}(M^{-1}g, (M^{-1}H)M^{-1}g, \ldots, (M^{-1}H)^k M^{-1}g)
\]

of \( M^{-1}H \) on \( M^{-1}g \) to partially reduce \( H \) to the tridiagonal form

\[
T_k = Q_k^T HQ_k = \begin{bmatrix}
\delta_0 & \gamma_1 & & \\
\gamma_1 & \delta_1 & \gamma_2 & \\
& \ddots & \ddots & \ddots \\
& & \gamma_{k-1} & \delta_{k-1} & \gamma_k \\
& & & \gamma_k & \delta_k
\end{bmatrix}, \quad (3.1)
\]

where \( Q_k^T MQ_k = I_{k+1} \) [6, Algorithm 4.2], assuming

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

Compactly, the process can be expressed by the relation

\[
HQ_k - MQ_k T_k = \gamma_{k+1} M q_{k+1} e_{k+1}^T
\]

with \( \gamma_0 = \|M^{-1}g\|_2 \), \( Q_k e_1 = q_0 : = \gamma_0^{-1}(M^{-1}g) \). Afterwards, we have the following reduced trust-region subproblem

\[
\min_{\|h\|_2 \leq \Delta} \tilde{f}(h) \quad \text{with} \quad \tilde{f}(h) := \frac{1}{2} h^T T_k h + \gamma_0 h^T e_1. \quad (3.3)
\]

Let \( h_k \) be the minimizer of (3.3). It can be verified that the vector \( s_k = Q_k h_k \in \mathcal{K}_k(M^{-1}H, M^{-1}g) \)

is the minimizer of

\[
\min_{s \in \mathcal{K}_k(M^{-1}H, M^{-1}g)} f(s), \quad (3.4)
\]

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

Generically, \( \dim \mathcal{K}_k(M^{-1}H, M^{-1}g) \) strictly increases by 1 as \( k \) increases by 1 and thus often \( \dim \mathcal{K}_k(M^{-1}H, M^{-1}g) = k + 1 \) until \( k = n - 1 \). But it can happen that \( \dim \mathcal{K}_k(M^{-1}H, M^{-1}g) \) 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 \( \ell \) be the smallest nonnegative integer such that

\[
\dim \mathcal{K}_\ell(M^{-1}H, M^{-1}g) = \dim \mathcal{K}_{\ell+1}(M^{-1}H, M^{-1}g) = \ell + 1. \quad (3.5)
\]

This is reflected by \( \gamma_{\ell+1} = 0 \) while \( \gamma_k \neq 0 \) for all \( 0 \leq k \leq \ell \). In such a case, \( HQ_\ell = MQ_\ell T_\ell \).

Two special cases for \( M = I_n \) are worth mentioning:

---

\(^2\)We adopt a notation convention that is consistent with the one used in [6]. That is to use \( \mathcal{K}_k \) for the Krylov 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.
1. the case \( g = 0 \). Both TLTRS and GLTR reduce to the classical Lanczos method for finding the smallest eigenpair of \( H \);

2. the case when \( H \) is positive definite and \( \Delta \geq \|H^{-1}g\|_2 \). Both TLTRS and GLTR are equivalent to CG for solving the linear system \( Hs = -g \) (see subsection 4.2 for detail).

Since CG is closely related to using the Lanczos process to partially reduce the coefficient matrix \( H \) to the tridiagonal form \([4, p.307]\), both CG and the Lanczos method for the symmetric eigenvalue problem can be thought of as Lanczos process-based methods. In view of this, we may say that GLTR and TLTRS methods lie between the Lanczos-based method for the linear system and that for the symmetric eigenvalue problem.

**Remark 3.1.** 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

\[
H \leftarrow M^{-1/2}HM^{-1/2}, \quad g \leftarrow M^{-1/2}g.
\]

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

\[
HQ_k - Q_kT_k = \gamma_{k+1}q_{k+1}e_{k+1}^\top, \quad \gamma_0 = \|g\|_2, \quad Q_ke_1 = g/\gamma_0.
\]  

(3.6)

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

\[
Q^\top HQ = \begin{bmatrix}
\delta_0 & \gamma_1 & & \\
\gamma_1 & \delta_1 & \ddots & \\
& \ddots & \ddots & \gamma_k \\
& & \gamma_k & \delta_k \\
& & & \gamma_{k+1} \\
& & & & \ddots \\
& & & & \gamma_\ell \\
\gamma_{k+1}e_{k+1}^\top & & & T_k \\
\end{bmatrix} =: T.
\]  

(3.7)

Denote the eigenvalues of \( T_k \) by \( \sigma_i^{(k)} \) arranged in the increasing order:

\[
\sigma_1^{(k)} \leq \sigma_2^{(k)} \leq \cdots \leq \sigma_{k+1}^{(k)}.
\]
where the superscript \(^{(k)}\) indicates the association of \(\sigma_i^{(k)}\) with the order of the Krylov subspace \(\mathcal{K}_k(H, g)\).

Associated with every Lanczos step \(k\) before breakdown is the corresponding TRS (3.3). Let \(h_k\) and \(\lambda_k\) be the solution of (3.3) and the Lagrangian multiplier for it, respectively, and set \(s_k = Q_k h_k\). With these settings, the following lemma follows.

**Lemma 3.1.** We have

(i) for any \(k = 0, 1, \ldots, \ell\),

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

(ii) for \(0 \leq j \leq k \leq \ell\),

\[ \sigma_i^{(k)} \leq \sigma_i^{(j)}, \quad \text{for} \quad i = 1, 2, \ldots, j+1; \]

(iii) in the nondegenerate case, \(s_\ell = s_{\text{opt}}\) and \(\lambda_\ell = \lambda_{\text{opt}}\).

**Proof.** The inequalities in items (i) and (ii) are straightforward consequences of Cauchy interlacing inequalities [18].

Item (iii) for the case \(g \not\perp \mathcal{E}_1\) has been proved in [6, Theorem 5.7]. We consider the special scenario:

\[ g \perp \mathcal{E}_1 \quad \text{but} \quad \|(H - \theta_1 I_n)^{\dagger} g\|_2 > \Delta. \]

Define \(\rho(\lambda) = \|(H + \lambda I_n)^{\dagger} g\|_2\), and it is true by \(g \perp \mathcal{E}_1\) that \(\rho(\lambda)\) is a continuous and nonincreasing function of \(\lambda \in (- \theta_{p+1}, +\infty)\). Also note from \(Q^{\dagger} g = \gamma_0 e_1\) and (3.7) with \(\gamma_{\ell+1} = 0\) that for \(\lambda > - \theta_{p+1}\)

\[ \rho(\lambda) = \|(T^{\dagger} + \lambda I_n)^{\dagger} e_1\|_2^{2} = \gamma_0 \|(T + \lambda I_n)^{\dagger} e_1\|_2^{2} = \gamma_0 \|(T_\ell + \lambda I_{\ell+1})^{\dagger} e_1\|_2^{2}, \]

implying that \(\gamma_0 \|(T_\ell + \lambda I_{\ell+1})^{\dagger} e_1\|_2\) is also a continuous and nonincreasing function of \(\lambda > - \theta_{p+1}\). Thus, from

\[ \rho(\lambda) = \gamma_0 \|(T_\ell + \lambda I_{\ell+1})^{\dagger} e_1\|_2 \leq \Delta < \rho(- \theta_1), \]

we know \(\lambda_\ell > - \theta_1\), i.e., \(H + \lambda I_n\) is positive definite. Moreover, by [6, Theorem 5.1], \((H + \lambda I_n) s_\ell = -g\), which according to Lemma 2.1 and the uniqueness in the nondegenerate case leads to (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 \(s_{\text{opt}}\) is of the form (2.3) with \(\tau > 0\). As the Lanczos process starting from \(g\) cannot extract any information out of the eigenspace \(\mathcal{E}_1\), the approximate solution \(s_k = Q_k h_k\) does not contain the component of \(\tau u\) for any \(u \in \mathcal{E}_1\), even for \(k = \ell\). In other words, 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 also discussed in [6, 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 [14], which claims that the Lagrangian multipliers \(\lambda_k\) monotonically increases with \(k\).
Theorem 3.1 ([14]). The sequence \( \{\lambda_k\}_{k=0}^\ell \) of Lagrangian multipliers associated with (3.3) is monotonically nondecreasing in \( k \).

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

**Proposition 3.1.** Let \( 0 \leq k \leq \ell \), then

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

(ii) in the nondegenerate case, \( 0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_\ell = \lambda_{\text{opt}} \).

4 Convergence analysis for TLTRS

Throughout this section, we assume that (1.1) is nondegenerate, unless otherwise explicitly stated differently. We will discuss the convergence of TLTRS and GLTR for the two cases \( \lambda_{\text{opt}} = 0 \) and \( \lambda_{\text{opt}} > 0 \), separately.

The Chebyshev polynomials will show up in our convergence analysis. The \( k \)th Chebyshev polynomial of the first kind \( \mathcal{T}_k(t) \) is

\[
\mathcal{T}_k(t) = \cos(k \arccos t) = \frac{1}{2} \left[ (t + \sqrt{t^2 - 1})^k + (t - \sqrt{t^2 - 1})^{-k} \right] \quad \text{for } |t| \geq 1.
\]

It frequently shows up in numerical analysis and computations because of its numerous nice properties, for example \( |\mathcal{T}_k(t)| \leq 1 \) for \( |t| \leq 1 \) and \( |\mathcal{T}_k(t)| \) grows extremely fast\(^3\) for \( |t| > 1 \). We will also need [11]

\[
|\mathcal{T}_k \left( \frac{1+t}{1-t} \right)| = |\mathcal{T}_k \left( \frac{t+1}{t-1} \right)| = \frac{1}{2} \left[ \Gamma_t^k + \Gamma_t^{-k} \right] \quad \text{for } 1 \neq t > 0,
\]

where

\[
\Gamma_t := \frac{\sqrt{t} + 1}{|\sqrt{t} - 1|} \quad \text{for } t > 0. \tag{4.1}
\]

4.1 Convergence when \( \lambda_{\text{opt}} = 0 \)

In this case, \( H \) is positive definite, and moreover \( \|H^{-1}g\|_2 \leq \Delta \), implying that (1.1) is equivalent to the linear system: \( Hs_{\text{opt}} = -g \), and

\[
f(s) = \frac{1}{2} (s_{\text{opt}} - s)^\top H (s_{\text{opt}} - s).
\]

Moreover, by Lemma 3.1(iii) and Proposition 3.1, we know that \( \lambda_k = 0 \) for all \( k = 0, 1, \ldots, \ell \), which implies that each TRS (3.3) is equivalent to the linear system: \( T_\ell h_k = \)

\(^3\)In fact, a result due to Chebyshev himself says that if \( p(t) \) is a polynomial of degree no bigger than \( k \) and \( |p(t)| \leq 1 \) for \( -1 \leq t \leq 1 \), then \( |p(t)| \leq |\mathcal{T}_k(t)| \) for any \( t \) outside \([-1, 1]\) [2, p.65].
−γ0e1, and TLTRS turns out to be the Full Orthogonalization Method (FOM) [24, Algorithm 6.4]. Indeed,

\[ \mathbf{s}_k = \mathbf{Q}_k \mathbf{h}_k = \arg \min_{\mathbf{s} \in \mathcal{X}_k(H, g)} \frac{1}{2} (\mathbf{s}_{\text{opt}} - \mathbf{s})^\top H (\mathbf{s}_{\text{opt}} - \mathbf{s}), \]

the same as the one obtained from CG [24, Section 6.7] on the linear system \( H \mathbf{s}_{\text{opt}} = -g \). In other words, in this case, GLTR [6, Algorithm 5.1] will never hit the boundary of \( \| \mathbf{s} \|_2 \leq \Delta \) and thereby, the approximation \( \mathbf{s}_k \) from either TLTRS or GLTR is the same as that from the CG iteration for \( H \mathbf{s}_{\text{opt}} = -g \). Consequently, the standard convergence theory [24, Section 6.11.3] for CG applies for this situation. In particular, we have

\[ \| \mathbf{s}_{\text{opt}} - \mathbf{s}_k \|_H \leq \frac{1}{\mathcal{F}_k + 1 ((\kappa + 1)/(\kappa - 1))} = 2 [\Gamma_k^{k+1} + \Gamma_k^{-(k+1)}]^{-1}, \]

where \( \Gamma_k \) is defined by (4.1), and

\[ \kappa \equiv \kappa(H) = \frac{\theta_n}{\theta_1} \]

is the spectral condition number of \( H \). In terms of the spectral norm, (4.2) implies

\[ \frac{\| \mathbf{s}_{\text{opt}} - \mathbf{s}_k \|_2}{\| \mathbf{s}_{\text{opt}} \|_2} \leq \frac{\sqrt{\kappa}}{\mathcal{F}_k + 1 ((\kappa + 1)/(\kappa - 1))}. \]

Other discussions of the truncated conjugate gradient method for TRS in the case when \( H \) is positive definite can be found in [31]

### 4.2 Convergence when \( \lambda_{\text{opt}} > 0 \)

If \( \lambda_{\text{opt}} > 0 \), then \( \| \mathbf{s}_{\text{opt}} \|_2 = \Delta \). This is the case when \( H \) is indefinite or positive definite but \( \| H^{-1} g \|_2 > \Delta \). The following simple example demonstrates a case where \( \lambda_{\text{opt}} > 0 \) but \( \lambda_k = 0 \) for some \( 0 \leq k \leq \ell \), i.e., \( \mathbf{h}_k = -\gamma_0 T_k^{-1} e_1 \) is the solution to the projected TRS (3.3).

**Example 4.1.** Consider TRS with

\[ H = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \quad g = e_1 \in \mathbb{R}^2 \text{ and } \Delta = 1. \]

It can be verified that \( \lambda_{\text{opt}} \approx 0.1701 \) and \( \mathbf{s}_{\text{opt}} \approx [-0.7602, 0.6497]^\top \), but \( \lambda_0 = 0 \) and \( \mathbf{s}_0 = -e_1/2 \).

Even though \( \lambda_k = 0 \) might happen in the first stages of TLTRS, eventually \( \lambda_k > 0 \) as \( k \) increases, and thereby \( \| \mathbf{s}_k \|_2 = \Delta \). This also means that GLTR will eventually encounter the boundary of \( \| \mathbf{s} \|_2 \leq \Delta \), and proceeds as TLTRS afterwards. For that reason, in what follows, we analyze the errors

\[ \| \mathbf{s}_{\text{opt}} - \mathbf{s}_k \|_2 \quad \text{and} \quad | f(\mathbf{s}_k) - f(\mathbf{s}_{\text{opt}}) | \]

under the assumption \( \| \mathbf{s}_k \|_2 = \Delta \). Set

\[ H_{\text{opt}} := H + \lambda_{\text{opt}} I_n \]

which is positive definite since it is in the nondegenerate case.
4.2.1 The optimal polynomial

Denote by \( P_k \) the set of all polynomials of degree no higher than \( k \). Then the approximate solution \( s_k = Q_k h_k \in K_k(H, g) \) can be expressed as

\[
s_k = \psi_k(H)g = U\psi_k(\Theta)U^Tg = U\psi_k(\Theta)a = \sum_{i=1}^{n} \psi_k(\theta_i)a_i u_i,
\]

where \( a = U^Tg \) and the optimal polynomial \( \psi_k \in P_k \) is given by

\[
\psi_k = \arg \min_{\psi \in P_k} \|\psi(H)g\|_2 = \Delta_f(\psi(H)g).
\] (4.4)

To gain an alternative view of the optimization in (3.4), we now analyze the right-hand side of (4.4). Let

\[
\psi_k(\theta) = \sum_{i=0}^{k} \hat{p}_i \theta^i = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k+1}]_\theta = \hat{p}_\theta
\]

and

\[
V_{k+1,n} = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
\theta_1 & \theta_2 & \cdots & \theta_n \\
\vdots & \vdots & \ddots & \vdots \\
\theta_1^k & \theta_2^k & \cdots & \theta_n^k
\end{bmatrix}
= [c_k(\theta_1), \ldots, c_k(\theta_n)] \in \mathbb{R}^{(k+1) \times n}.
\]

Any \( x \in K_k(H, g) \) takes the form

\[
x = \hat{\psi}(H)g = U\hat{\psi}(\Theta)a = UD_a V_{k+1,n}^T p,
\]

where \( D_a = \text{diag}(a_1, a_2, \ldots, a_n) \) and \( p \in \mathbb{R}^{k+1} \) is the coefficient vector of \( \psi \). Consequently, the coefficient vector \( \hat{p} \) corresponding to \( \psi_k \) is given by

\[
\hat{p} = \arg \min_{\|D_a V_{k+1,n}^T p\|_2 = \Delta} \frac{1}{2} p^T (V_{k+1,n} D_a \Theta D_a V_{k+1,n}^T) p + p^T V_{k+1,n} D_a a.
\]

By the Lagrangian multiplier theory, there is \( \varrho_k \in \mathbb{R} \) such that

\[
[V_{k+1,n} D_a (\Theta + \varrho_k I_n) D_a V_{k+1,n}^T] \hat{p} = -V_{k+1,n} D_a a \quad \text{and} \quad \|D_a V_{k+1,n}^T \hat{p}\|_2 = \Delta.
\]

This alternative characterization for the optimal \( \psi_k \) doesn’t lead to a simple convergence analysis, however, as oppose to the ones in [11, 10] for analyzing CG, the minimal residual method, and the generalized minimal residual method. The main difficulty is the emergence of the Lagrangian multiplier \( \varrho_k \) or equivalently, the equality constraint \( \|D_a V_{k+1,n}^T \hat{p}\|_2 = \Delta \). 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 \( \|s_{opt}\|_2 = \Delta \) and on approximations \( s \) with \( \|s\|_2 = \Delta \). We first present a general framework to bound the errors of \( f(s_k) - f(s_{opt}) \) and \( \|s_k - s_{opt}\|_2 \) in terms of any nonzero \( \tilde{s} \in \mathcal{K}_k(H, g) \). Later, this framework will be realized for \( \tilde{s} \) constructed from certain sub-optimal polynomial \( \tilde{\psi}_k \) as opposed to the optimal one given by (4.4) for the purpose of established error bounds for TLTRS solutions.

\[ \tilde{s} = \tilde{\psi}_k(H)g \in \mathbb{R}^n \]  

(4.5)

\textbf{Theorem 4.1.} Suppose \( \|s_{opt}\|_2 = \Delta \) and \( s_k \) is the \( k \)th (\( k \leq \ell \)) approximation of TLTRS satisfying \( \|s_k\|_2 = \Delta \). Then for any nonzero \( \tilde{s} \in \mathcal{K}_k(H, g) \), we have

\[
0 \leq f(s_k) - f(s_{opt}) \leq 2\|H_{opt}\|_2\|\tilde{s} - s_{opt}\|_2^2, \quad \text{and} \quad \|s_k - s_{opt}\|_2 \leq 2\sqrt{\kappa}\|\tilde{s} - s_{opt}\|_2, \\
\]

(4.6)

(4.7)

where \( H_{opt} \) is given by (4.3) and

\[ \kappa \equiv \kappa(H_{opt}) = \frac{\theta_n + \lambda_{opt}}{\theta_1 + \lambda_{opt}} \]

is the condition number of \( H_{opt} \).

**Proof.** First, we have \( \|\tilde{s}\|_2 - \Delta \|\tilde{s}\|_2 = \|\tilde{s}\|_2 - \|s_{opt}\|_2 \leq \|\tilde{s} - s_{opt}\|_2 \) which leads to

\[
1 - \frac{\Delta}{\|s\|_2} \leq \frac{\|\tilde{s} - s_{opt}\|_2}{\|\tilde{s}\|_2}. \\
\]

(4.8)

Let \( r = \tilde{s} - s_{opt} \) where \( \tilde{s} = \frac{\tilde{s}}{\|\tilde{s}\|_2} \Delta \). We then have

\[
\|r\|_2 = \|s_{opt} - \tilde{s}\|_2 \leq \|s_{opt} - \tilde{s}\|_2 + \|\tilde{s} - \tilde{s}\|_2 \\
\leq \|s_{opt} - \tilde{s}\|_2 + \|\tilde{s} - \Delta \cdot \tilde{s} \|_2 \\
= \|s_{opt} - \tilde{s}\|_2 + \|\tilde{s}\|_2 \times 1 - \frac{\Delta}{\|\tilde{s}\|_2} \\
\leq 2\|s_{opt} - \tilde{s}\|_2, \\
\]

(4.9)

where the last inequality is obtained by using (4.8). Moreover, since for any \( 0 \leq i \leq \ell - 1 \),

\[
f(s_i) = \min_{s \in \mathcal{K}_i(H, g)} f(s) \geq \min_{s \in \mathcal{K}_{i+1}(H, g)} f(s) = f(s_{i+1}) \geq \min_{s \in \mathcal{K}_{i}(H, g)} f(s) = f(s_{opt}), \\
\]

(4.10)

we have

\[
0 \leq f(s_k) - f(s_{opt}) \\
\]

we have
\[
\begin{align*}
\leq f(\tilde{s}) - f(s_{\text{opt}}) &= \frac{r^\top Hr}{2} + r^\top (Hs_{\text{opt}} + g) \\
&= \frac{r^\top Hr}{2} - \lambda_{\text{opt}} r^\top s_{\text{opt}} \\
&= \frac{r^\top (H + \lambda_{\text{opt}} I)r}{2} \\
&\leq \frac{\|H_{\text{opt}}\|_2 \|r\|_2^2}{2} \\
&\leq 2\|H_{\text{opt}}\|_2 \|s_{\text{opt}} - \tilde{s}\|_2^2,
\end{align*}
\] (4.11)

where for obtaining (4.11) we used
\[
\Delta^2 = \|\tilde{s}\|^2_2 = \|s_{\text{opt}}\|^2_2 + \|r\|^2_2 + 2r^\top s_{\text{opt}}
\]
to get \(r^\top s_{\text{opt}} = -\|r\|^2_2/2 = -r^\top r/2\), and used (4.9) for getting (4.12).

For (4.7), we define
\[
f_{\text{opt}}(s) := \frac{1}{2} s^\top H_{\text{opt}} s + s^\top g = f(s) + \lambda_{\text{opt}} \|s\|^2_2.
\]
Then by noting that \(\nabla f_{\text{opt}}(s_{\text{opt}}) = H_{\text{opt}} s_{\text{opt}} + g = 0\), we have for any \(s\),
\[
f_{\text{opt}}(s) = f_{\text{opt}}(s_{\text{opt}}) + \frac{1}{2} (s - s_{\text{opt}})^\top H_{\text{opt}} (s - s_{\text{opt}}),
\]
and thus,
\[
f_{\text{opt}}(s) - f_{\text{opt}}(s_{\text{opt}}) \geq \frac{1}{2} (\theta_1 + \lambda_{\text{opt}}) \|s - s_{\text{opt}}\|^2_2. \tag{4.13}
\]
Furthermore, if \(\|s\|_2 = \Delta\), then
\[
f_{\text{opt}}(s) - f_{\text{opt}}(s_{\text{opt}}) = \left[ f(s) + \lambda_{\text{opt}} \|s\|^2_2 \right] - \left[ f(s_{\text{opt}}) + \lambda_{\text{opt}} \|s_{\text{opt}}\|^2_2 \right] = f(s) - f(s_{\text{opt}})
\]
since \(\|s_{\text{opt}}\|_2 = \Delta\) also. Consequently, for \(s_k\), by (4.6) and (4.13), we have
\[
\frac{1}{2} (\theta_1 + \lambda_{\text{opt}}) \|s_k - s_{\text{opt}}\|^2_2 \leq f(s_k) - f(s_{\text{opt}}) \leq 2\|H_{\text{opt}}\|_2 \|s_{\text{opt}} - \tilde{s}\|^2_2,
\]
which yields (4.7).

Next, we will discuss three sub-optimal polynomials \(\tilde{\psi}_k \in P_k\) to realize \(\tilde{s}\) by (4.5).

CG polynomials. Observe that \(s_{\text{opt}}\) is indeed the solution to the linear system
\[
(H + \lambda_{\text{opt}} I_n) s = H_{\text{opt}} s = -g,
\] (4.14)
where \( H_{\text{opt}} = H + \lambda_{\text{opt}}I_n \) is symmetric and positive definite, and that \( \mathcal{K}_k(H, g) = \mathcal{K}_k(H_{\text{opt}}, g) \). Apply CG to the linear system (4.14). The \( k \)th CG solution can be expressed as \( \tilde{s}^{cg}_k = \psi^{cg}_k(H)g \), where \( \psi^{cg}_k \in \mathbb{P}_k \). This \( \psi^{cg}_k \) is our first sub-optimal polynomial. The classical convergence result of CG says [4, 7, 11, 10, 24]

\[
\frac{\|\tilde{s}^{cg}_k - s_{\text{opt}}\|_{H_{\text{opt}}}}{\|s_{\text{opt}}\|_{H_{\text{opt}}}} \leq \frac{1}{\mathcal{F}_{k+1}(\eta)} = 2\left[ \Gamma_{\kappa+1}^{-1} + \Gamma_{\kappa}^{-1}(k+1) \right]^{-1}
\]

where

\[
\eta = \frac{\kappa + 1}{\kappa - 1} \quad \text{with} \quad \kappa = \kappa(H_{\text{opt}}) = \frac{\theta_1 + \lambda_{\text{opt}}}{\theta_1 + \lambda_{\text{opt}}}
\]

Thus,

\[
\|\tilde{s}^{cg}_k - s_{\text{opt}}\|_2 \leq \frac{\sqrt{\kappa} \Delta}{\mathcal{F}_{k+1}(\eta)}.
\]

Now with \( \tilde{s} = \tilde{s}^{cg}_k \), by Theorem 4.1, we have

\[
0 \leq f(s_k) - f(s_{\text{opt}}) \leq 2\|H_{\text{opt}}\|_2 \left( \frac{\sqrt{\kappa} \Delta}{\mathcal{F}_{k+1}(\eta)} \right)^2,
\]

\[
\|s_{\text{opt}} - s_k\|_2 \leq 2\sqrt{\kappa} \frac{\sqrt{\kappa} \Delta}{\mathcal{F}_{k+1}(\eta)} = \frac{2\kappa \Delta}{\mathcal{F}_{k+1}(\eta)}.
\]

**MINRES polynomials.** The second choice for \( \tilde{s} \in \mathcal{K}_k(H, g) \) is the one obtained from applying the minimal residual method (MINRES) to solve (4.14). The \( k \)th MINRES solution can be expressed as \( \tilde{s}^{mr}_k = \psi^{mr}_k(H)g \), where \( \psi^{mr}_k \in \mathbb{P}_k \) is also a sub-optimal polynomial. We remark that even though \( H_{\text{opt}} \) is symmetric and positive definite, \( \tilde{s}^{mr}_k \) may possibly deliver a more accurate approximation than \( \tilde{s}^{cg}_k \) in terms of the objective value. Basing on the convergence result of MINRES [24, Proposition 6.32], we have for \( \tilde{s} = \tilde{s}^{mr}_k \) that

\[
\|\tilde{s}^{mr}_k - s_{\text{opt}}\|_2 \leq \frac{\|g\|_2}{\mathcal{F}_{k+1}(\eta)} = 2\|g\|_2 \left[ \Gamma_{\kappa+1}^{-1} + \Gamma_{\kappa}^{-1}(k+1) \right]^{-1}.
\]

Now with \( \tilde{s} = \tilde{s}^{mr}_k \), by Theorem 4.1, we have

\[
0 \leq f(s_k) - f(s_{\text{opt}}) \leq 2\|H_{\text{opt}}\|_2 \left( \frac{\|g\|_2}{\mathcal{F}_{k+1}(\eta)} \right)^2,
\]

\[
\|s_{\text{opt}} - s_k\|_2 \leq 2\sqrt{\kappa} \frac{\|g\|_2}{\mathcal{F}_{k+1}(\eta)}.
\]

We note that \( \eta \) is the dominant factor in both upper bounds from the CG and MINRES polynomials. The bigger \( \eta \) is, the faster TLTRS converges.

**Best polynomials for approximating \( \frac{1}{x^\gamma} \).** Lastly, we discuss yet another sub-optimal polynomial. Note that \( g = Ua = \sum_{i=1}^n u_i a_i \), and we have for any \( x \in \mathcal{K}_k(H, g) \)

\[
x = \psi(H)g = U\psi(\Theta)a = \sum_{i=1}^n \psi(\theta_i)a_i u_i \quad \text{for some} \quad \psi \in \mathbb{P}_k, \quad \text{and} \quad (4.18)
\]

13
\[ s_{\text{opt}} = -(H + \lambda_{\text{opt}} I_n)^{-1} g = -U(\Theta + \lambda_{\text{opt}} I_n)^{-1} a = -\sum_{i=1}^{n} \frac{a_i}{\theta_i + \lambda_{\text{opt}}} u_i. \] (4.19)

By comparing (4.18) and (4.19), we define a sub-optimal polynomial \( \hat{\psi}_k^{\text{ra}} \in \mathbb{P}_k \) as the solution to the minimax approximation problem:

\[ \hat{\psi}_k^{\text{ra}} = \arg \min_{\psi \in \mathbb{P}_k} \max_{\theta_1 \leq \theta \leq \theta_n} \left| \psi(\theta) - \frac{1}{\theta + \lambda_{\text{opt}}} \right|. \] (4.20)

In the other word, \( \hat{\psi}_k^{\text{ra}} \) is the best polynomial of approximation to the rational function \( \frac{1}{\theta + \lambda_{\text{opt}}} \) 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 noting (4.15):

\[ \eta = \frac{x + 1}{x - 1} = \frac{\theta_1 + \theta_n + 2\lambda_{\text{opt}}}{\theta_n - \theta_1} > 1, \]

we have

\[
\begin{aligned}
&\min_{\psi \in \mathbb{P}_k} \max_{\theta_1 \leq \theta \leq \theta_n} \left| \psi(\theta) - \frac{1}{\theta + \lambda_{\text{opt}}} \right| \\
&= \min_{\psi \in \mathbb{P}_k} \max_{-1 \leq x \leq 1} \left| \psi(\theta(x)) - \frac{2}{(\theta_n - \theta_1)(x - \frac{\theta_1 + \theta_n + 2\lambda_{\text{opt}}}{\theta_n - \theta_1})} \right| \\
&= \frac{2}{\theta_n - \theta_1} \times \min_{\psi \in \mathbb{P}_k} \max_{-1 \leq x \leq 1} \left| \psi(x) - \frac{1}{x - \eta} \right|,
\end{aligned}
\]

which implies that

\[
\begin{aligned}
&\max_{\theta_1 \leq \theta \leq \theta_n} \left| \hat{\psi}_k^{\text{ra}}(\theta) - \frac{1}{\theta + \lambda_{\text{opt}}} \right| \\
&= \frac{2}{\theta_n - \theta_1} \times \min_{\psi \in \mathbb{P}_k} \max_{-1 \leq x \leq 1} \left| \psi(x) - \frac{1}{x - \eta} \right|, \\
&= \epsilon_k^\text{ra}(\eta),
\end{aligned}
\] (4.21)

where \( \epsilon_k^\text{ra}(\eta) \) 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 \( \epsilon_k^\text{ra}(\eta) \) with respect to \( k \) and \( \eta \), 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 (see [15, Section 4.3] for more detailed information).

**Lemma 4.1** (Bernstein [1]). For \( \eta > 1 \), suppose \( p_k(x) \in \mathbb{P}_k \) is the best approximating polynomial of \( \frac{1}{x - \eta} \) in \([-1, 1]\), then

\[
\frac{1}{x - \eta} - p_k(x) = \left( \frac{\eta + \sqrt{\eta^2 - 1}}{\eta^2 - 1} \right)^k \cos(k\alpha + \beta),
\]
where \( x = \cos \alpha \) and \( \eta_x - 1 = \cos \beta \), and moreover,

\[
\epsilon^r_k(\eta) = \min_{\psi \in \mathbb{P}_k} \max_{1 \leq i \leq 1} \left| \psi(x) - \frac{1}{x - \eta} \right| = \frac{(\eta + \sqrt{\eta^2 - 1})^{-k}}{\eta^2 - 1}. \tag{4.22}
\]

**Remark 4.1.** It is noted that \( \eta + \sqrt{\eta^2 - 1} > \eta > 1 \) since \( \eta > 1 \), and for \( \eta \) given by (4.15),

\[
\eta + \sqrt{\eta^2 - 1} = \Gamma_x.
\]

Therefore, \( \epsilon^r_k(\eta) \) converges linearly to zero with the linear factor \( \Gamma_x^{-1} \) as \( k \) increases.

Now we can establish error bounds for TLTRS solutions in terms of \( \epsilon^r_k(\eta) \). The corresponding estimates for \( \|s_k - s_\text{opt}\|_2 \) and \( f(s_k) - f(s_\text{opt}) \) also reflects the behavior of TLTRS characterized by the number of Lanczos step \( k \) and the parameter \( \eta \) as we will see from the numerical examples in section 5.

Let \( \tilde{\psi}_k \) be defined by (4.20), and set \( \tilde{s}_k^r = \tilde{\psi}_k^r(H)g \). Note by (4.21) that

\[
\|\tilde{s}_k^r - s_\text{opt}\|_2 = \left\| U \left( \tilde{\psi}_k^r(\theta) - (\theta + \lambda_\text{opt})^{-1} \right) a \right\|_2
\]

\[
\leq \|g\|_2 \times \max_{\theta = \theta_1, \ldots, \theta_n} \left| \tilde{\psi}_k^r(\theta) - \frac{1}{\theta + \lambda_\text{opt}} \right|
\]

\[
\leq \|g\|_2 \times \max_{\theta_1 \leq \theta \leq \theta_n} \left| \tilde{\psi}_k^r(\theta) - \frac{1}{\theta + \lambda_\text{opt}} \right|
\]

\[
= \frac{2\|g\|_2}{\theta_n - \theta_1} \epsilon^r_k(\eta).
\]

Now with \( \tilde{s} = \tilde{s}_k^r \), by Theorem 4.1, we have

\[
0 \leq f(s_k) - f(s_\text{opt}) \leq 2\|H_\text{opt}\|_2 \left( \frac{2\|g\|_2}{\theta_n - \theta_1} \epsilon^r_k(\eta) \right)^2, \tag{4.23a}
\]

\[
\|s_\text{opt} - s_k\|_2 \leq 2\sqrt{\epsilon^r_k(\eta)} = \frac{4\sqrt{\epsilon^r_k(\eta)}}{\theta_n - \theta_1} \epsilon^r_k(\eta). \tag{4.23b}
\]

Summarizing the results for all three sub-optimal solutions above yields the item (ii) of the following theorem.

**Theorem 4.2.** Let the sequence \( \{s_k\}_{k=0}^\ell \) be generated by TLTRS for TRS (1.1).

(i) The sequence \( \{f(s_k)\}_{k=0}^\ell \) is nonincreasing, and \( f(s_\ell) = f(s_\text{opt}) \) for the nondegenerate case, and

\[
f(s_\ell) + \frac{\tau^2 \theta_1}{2} \leq f(s_\text{opt}) \leq f(s_\ell)
\]

for the degenerate case, where \( \tau^2 = \Delta^2 - \| (H - \theta_1 I_n)^\dagger g \|_2^2 > 0 \) and \( \theta_1 \leq 0 \);

(ii) For the nondegenerate case, if \( \|s_\text{opt}\|_2 \leq \|s_k\|_2 = \Delta \) for some \( 0 \leq k \leq \ell \), then

\[
0 \leq f(s_k) - f(s_\text{opt}) \leq 2\|H_\text{opt}\|_2 \zeta_{k}^2,
\]

\[
(4.25a)
\]

15
where $H_{\text{opt}} = H + \lambda_{\text{opt}} I_n$, and

$$
\kappa = \theta_n + \lambda_{\text{opt}} \overline{\zeta_k} = \min \left\{ \frac{\sqrt{\Delta}}{\mathcal{R}_{k+1}(\eta)}, \frac{\|g\|_2}{\mathcal{R}_{k+1}(\eta)}, \frac{2\|g\|_2 \tau_{k}^{\eta}(\eta)}{\theta_n - \theta_1} \right\} .
$$

**(Proof.)** Based on our previous discussions, only the inequality (4.24) needs a proof. First, it is easy to see that $f(s_\ell)$ is an upper bound for $f(s_{\text{opt}})$ by (4.10). For the orthogonal matrix $Q := [Q, P] \in \mathbb{R}^{n \times n}$ satisfying (3.7), let

$$
Q^T s_{\text{opt}} = \begin{bmatrix} Q^T s_{\text{opt}} \\ P^T s_{\text{opt}} \end{bmatrix} = \begin{bmatrix} y \\ z \end{bmatrix}.
$$

From $Q^T H Q = T = \text{diag}(T_\ell, \tilde{T}_\ell)$, $T_\ell = T_{(1, \ell+1, 1, \ell+1)}$, $\tilde{T}_\ell = T_{(\ell+2, n, \ell+2, n)}$ and $Q^T g = \gamma_0 e_1$, we have

$$
f(s_{\text{opt}}) = \frac{1}{2} s_{\text{opt}}^T H s_{\text{opt}} + g^T s_{\text{opt}}
= \frac{1}{2} s_{\text{opt}}^T Q T Q^T s_{\text{opt}} + s_{\text{opt}}^T Q Q^T g
\geq \frac{1}{2} y^T T_\ell y + \gamma_0 y^T e_1 + \frac{1}{2} z^T \tilde{T}_\ell z,
$$

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

$$
\frac{1}{2} y^T T_\ell y + \gamma_0 y^T e_1 \geq \min_{\| h \|_2 \leq \Delta} f(Q_\ell h) = f(s_\ell).
$$

We next establish a lower bound for $\frac{1}{2} z^T \tilde{T}_\ell z$. In fact, we have

$$
\frac{1}{2} z^T \tilde{T}_\ell z \geq \frac{\theta_1 \| z \|_2^2}{2},
$$

since the condition $g \perp E_1$ for the degenerate case and the breakdown in the Lanczos process imply $\lambda_{\min}(T_\ell) = \theta_1$.

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

$$
E_1 = \mathcal{R}(U_1) \subseteq \mathcal{R}(P) \quad \text{and} \quad \mathcal{R}(Q_\ell) \subseteq \mathcal{R}(U_2),
$$

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

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

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

$$
P^T U_2 U_2^T Q_\ell = P^T (H - \theta_1 I_n)^\dagger Q_\ell(T_\ell - \theta_1 I_{\ell+1}).
$$

(4.28)
Since $Q_\ell^T P = 0$ and $Q_\ell^T U_1 = 0$ by (4.27), we know that
\[ P^T U_2 U_2^T Q_\ell = P^T (I_n - U_1 U_1^T) Q_\ell = P^T Q_\ell - P^T U_1 U_1^T Q_\ell = 0 \]
which, together with (4.28), lead to
\[ P^T (H - \theta_1 I_n) Q_\ell T_\ell = 0. \] (4.29)

Note $T_\ell - \theta_1 I_{\ell+1}$ is positive definite. Post-multiplying both sides of (4.29) by $(T_\ell - \theta_1 I_{\ell+1})^{-1} e_1$ and using $Q_\ell e_1 = g/\gamma_0$, we get $P^T (H - \theta_1 I_n)^T g = 0$, from which and (2.3) it follows that
\[ z = -P^T (H - \theta_1 I_n)^T g + \tau P^T u = \tau P^T u \Rightarrow \|z\|_2 = \tau, \]
as expected, where $u \in E_1$ is a unit vector.

It is noted that in all the three sub-optimal solutions, the factor $\eta$ 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 4.1 and Remark 4.1, we observe that both $\frac{1}{\gamma_{k+1}(\eta)}$ and $\epsilon_k(\eta)$ decay fast for big $\eta$. Note that
\[ \eta = 1 + \frac{2}{\kappa - 1} = 1 + \frac{2\theta_1 + \lambda_{\text{opt}}}{\theta_n - \theta_1}, \] (4.30)
and thus a big $\eta$ corresponds to the situation where TRS (1.1) is far away from the degenerate case (i.e., $\theta_1 > -\lambda_{\text{opt}}$). In other words, TLTRS converges fast when TRS is far away from the degenerate case, but needs more Lanczos steps when it is near the degenerate case. This conclusion is consistent with the numerical observations in [8] that “a Lanczos type process seems to be very effective when the problem is very nondegenerate”. This phenomenon is also reflected in our numerical examples in section 5. Theoretically, we have the following lower bound for $\eta$:

**Theorem 4.3.** If $\|s_{\text{opt}}\|_2 = \Delta$, then we have
\[ \eta \geq 1 + \frac{\|g\|_2 \cos \angle(g, E_1)}{\sqrt{\Delta^2 (\theta_n + \|H\|_2 + \|g\|_2/\Delta)^2 - \|g\|_2^2 \cos^2 \angle(g, E_2)}} \] (4.31)
where $\angle(g, E_1)$ and $\angle(g, E_2)$ stand for the angles from $\mathbb{R}(g)$ to $E_1 = \mathbb{R}(U_1)$ and $E_2 = \mathbb{R}(U_2)$, respectively.

**Proof.** Note that the factor $\eta$ defined in (4.15) can also be expressed by (4.30). The assertion is obvious for the degenerate case, i.e., $\lambda_{\text{opt}} = -\theta_1$ and $\eta = 1$, since $\cos \angle(g, E_1) = 0$.

Now consider the nondegenerate case. Then $\lambda_{\text{opt}} > -\theta_1$ and $H_{\text{opt}} = H + \lambda_{\text{opt}} I_n$ is positive definite. Thus $\max \{0, -\theta_1\} \leq \lambda_{\text{opt}}$. For this case, we also have $(H + \lambda_{\text{opt}} I_n) s_{\text{opt}} = -g$ and $\|s_{\text{opt}}\|_2 = \Delta$. Therefore
\[ \lambda_{\text{opt}} \Delta = \|\lambda_{\text{opt}} I_n s_{\text{opt}}\|_2 = \| -g - H s_{\text{opt}}\|_2 \leq \gamma_0 + \|H\|_2 \Delta. \]
Putting all together, we have
\[
\max\{0, -\theta_1\} \leq \lambda_{\text{opt}} \leq \|H\|_2 + \frac{7\theta_0}{\Delta}. \tag{4.32}
\]

Also, it holds that \(\cos \angle(\mathbf{g}, \mathbf{E}_i) = \frac{\|U_i^T \mathbf{g}\|_2}{\|\mathbf{g}\|_2}\) for \(i = 1, 2\), and with \(\|\mathbf{s}_{\text{opt}}\|_2 = \Delta\), or, equivalently,
\[
\Delta^2 = \sum_{i=1}^{n} \frac{(u_i^T \mathbf{g})^2}{(\theta_1 + \lambda_{\text{opt}})^2} = \sum_{i=1}^{p} \frac{(u_i^T \mathbf{g})^2}{(\theta_1 + \lambda_{\text{opt}})^2} + \sum_{i=p+1}^{n} \frac{(u_i^T \mathbf{g})^2}{(\theta_1 + \lambda_{\text{opt}})^2},
\]
which yields
\[
\Delta^2 \geq \left[ \frac{\cos^2 \angle(\mathbf{g}, \mathbf{E}_1)}{(\theta_1 + \lambda_{\text{opt}})^2} + \frac{\cos^2 \angle(\mathbf{g}, \mathbf{E}_2)}{(\theta_n + \lambda_{\text{opt}})^2} \right] \|\mathbf{g}\|_2^2.
\]

Multiply both sides by \((\theta_1 + \lambda_{\text{opt}})^2(\theta_n + \lambda_{\text{opt}})^2\) to get
\[
(\theta_1 + \lambda_{\text{opt}})^2 (\theta_n + \lambda_{\text{opt}})^2 \Delta^2 - \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_2) \geq \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_1)(\theta_n + \lambda_{\text{opt}})^2
\]
\[
\geq \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_1)(\theta_n - \theta_1)^2,
\]
where we have used \(\lambda_{\text{opt}} > -\theta_1\) for obtaining the last inequality. Therefore, by (4.32), we have
\[
\frac{\theta_1 + \lambda_{\text{opt}}}{\theta_n - \theta_1} \geq \frac{\|\mathbf{g}\|_2 \cos \angle(\mathbf{g}, \mathbf{E}_1)}{\sqrt{(\theta_n + \lambda_{\text{opt}})^2 \Delta^2 - \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_2)}}
\]
\[
\geq \sqrt{\Delta^2(\theta_n + \|H\|_2^2 + \|\mathbf{g}\|_2^2/\Delta)^2 - \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_2)}
\]
and the bound for \(\eta\) follows from (4.30). \(\square\)

The quantity within the square root sign in (4.31) can be expressed differently to show that it is always positive:
\[
\Delta^2(\theta_n + \|H\|_2^2)^2 + 2\Delta(\theta_n + \|H\|_2) + \|\mathbf{g}\|_2^2 \cos^2 \angle(\mathbf{g}, \mathbf{E}_1) > 0.
\]

**Remark 4.2.** We point out that the quantities \(\theta_1, \theta_n, \lambda_{\text{opt}}, \kappa\) and \(\eta\) involved in the upper bounds in (4.25a), (4.26), and (4.25b) are usually unknown. Fortunately, TLTRS is able to produce approximations to all these quantities: \(\sigma_1^{(k)} \approx \theta_1, \sigma_{k+1}^{(k)} \approx \theta_n, \lambda_k \approx \lambda_{\text{opt}}, \kappa = \frac{\theta_n + \lambda_{\text{opt}}}{\theta_1 + \lambda_{\text{opt}}} \approx \frac{\sigma_{k+1}^{(k)} + \lambda_k}{\sigma_1^{(k)} + \lambda_k} =: \kappa_k\) and \(\eta \approx 1 + 2\frac{\sigma_1^{(k)} + \lambda_k}{\sigma_{k+1}^{(k)} - \sigma_1^{(k)}} =: \eta_k\)
and these approximations are usually very good. Note that all the extra computation effort for these approximations is to compute the smallest and the largest eigenvalues of the tridiagonal \(T_k \in \mathbb{R}^{(k+1) \times (k+1)}\) given by (3.1). As \(k \ll n\) in general, obtaining these approximations can be very economical (with roughly \(O(k^2)\) flops). Consequently, practical estimates for \(f(\mathbf{s}_k) - f(\mathbf{s}_{\text{opt}})\) and \(\|\mathbf{s}_k - \mathbf{s}_{\text{opt}}\|_2\) instead of (4.25a) and (4.25b), respectively, are
\[
0 \leq f(\mathbf{s}_k) - f(\mathbf{s}_{\text{opt}}) \lesssim 2(\sigma_{k+1}^{(k)} + \lambda_k)\lambda_k^2 \quad \text{and} \quad \|\mathbf{s}_k - \mathbf{s}_{\text{opt}}\|_2 \lesssim 2\sqrt{\kappa_k} \lambda_k, \tag{4.33}
\]
where
\[
\chi_k = \min \left\{ \frac{\sqrt{\gamma_k}}{\mathcal{E}_{k+1}(\eta_k)} \frac{\|g\|_2}{\mathcal{E}_{k+1}(\eta_k)} \frac{2\|g\|_2 c_k^{\text{a}}(\eta_k)}{\sigma^{(k)}_{k+1} - \sigma^{(k)}_{1}} \right\}.
\]

We will see in our numerical examples in section 5 that these approximations can provide quite good upper bound estimates for \(f(s_k) - f(s_{\text{opt}})\) and \(\|s_k - s_{\text{opt}}\|_2\).

### 4.3 Convergence for the weighted-norm TRS

We next briefly show that our convergence results for \(M = I_n\) can be easily extended to TLTRS (with the preconditioned Lanczos process) for solving the weighted-norm TRS (1.2). Indeed, by defining
\[
v = M^{1/2}s, \quad \tilde{H} = M^{-1/2}HM^{-1/2} \quad \text{and} \quad \tilde{g} = M^{-1/2}g,
\]
we can rewrite the weighted-norm TRS (1.2) equivalently as
\[
\min_{\|v\|_2 \leq \Delta} \left\{ \tilde{f}(v) := \frac{1}{2} v^T \tilde{H} v + v^T \tilde{g} \right\}.
\]

Note the relations
\[
\tilde{f}(v) = f(s) \quad \text{and} \quad \|s\|_M = \|v\|_2.
\]

Let \(v_{\text{opt}}\) and \(s_{\text{opt}}\) be the optimal solutions to (4.34) and (1.2), respectively. It can be seen that their associated Lagrangian multipliers \(\lambda_{\text{opt}}\) are the same. Notice that
\[
\mathcal{K}_k(M^{-1}H, M^{-1}g) = M^{-1/2} \mathcal{K}_k(\tilde{H}, \tilde{g}),
\]
and therefore, with \(s = M^{-1/2}v\), it holds that
\[
s \in \mathcal{K}_k(M^{-1}H, M^{-1}g) \quad \text{with} \quad \|s\|_M \leq \Delta \quad \iff \quad v \in \mathcal{K}_k(\tilde{H}, \tilde{g}) \quad \text{with} \quad \|v\|_2 \leq \Delta.
\]

Recalling that the iterate \(s_k\) from TLTRS for (1.2) is defined by (3.4), we have \(s_k = M^{-1/2}v_k\), where
\[
v_k = \arg \min_{v \in \mathcal{K}_k(\tilde{H}, \tilde{g}), \|v\|_2 \leq \Delta} \tilde{f}(v).
\]

In other words, \(v_k\) is the iterate from TLTRS applying to (4.34) for which our main results in Theorem 4.2 are applicable.

**Theorem 4.4.** Let the sequence \(\{s_k\}_{k=0}^\ell\) be generated from TLTRS (with the preconditioned Lanczos process) for the weighted-norm TRS (1.2). Then

(i) The sequence \(\{f(s_k)\}_{k=0}^\ell\) is nonincreasing, and \(f(s_\ell) = f(s_{\text{opt}})\) for the nondegenerate case, and
\[
f(s_\ell) + \frac{\tilde{\tau}^2 \tilde{\theta}_1}{2} \leq f(s_{\text{opt}}) \leq f(s_\ell)
\]
for the degenerate case, where \(\tilde{\theta}_1 \leq \cdots \leq \tilde{\theta}_n\) are the ordered eigenvalues of \(\tilde{H} = M^{-1/2}HM^{-1/2}\), and \(\tilde{\tau}^2 = \Delta^2 - \|(H - \tilde{\theta}_1M)^Tg\|_M^2 > 0\) and \(\tilde{\theta}_1 \leq 0;\)
(ii) For the nondegenerate case, if \( \| \mathbf{s}_{\text{opt}} \|_M = \| \mathbf{s}_k \|_M = \Delta \) for some \( 0 \leq k \leq \ell \), then
\[
0 \leq f(\mathbf{s}_k) - f(\mathbf{s}_{\text{opt}}) \leq 2\| \hat{H}_{\text{opt}} \|_2 \hat{\zeta}_k^2,
\]
\[
\| \mathbf{s}_k - \mathbf{s}_{\text{opt}} \|_M \leq 2\sqrt{\kappa} \hat{\zeta}_k,
\]

where \( \hat{H}_{\text{opt}} = \hat{H} + \lambda_{\text{opt}} \mathbf{I}_n \), \( \hat{g} = M^{-1/2} \mathbf{g} \), and
\[
\hat{\zeta}_k = \min \left\{ \frac{\sqrt{\kappa} \Delta}{\mathcal{T}_{k+1}(\eta)}, \frac{\| \hat{\mathbf{g}} \|_2}{\mathcal{T}_{k+1}(\eta)}, \frac{2\| \hat{\mathbf{g}} \|_2 \epsilon^\beta(\eta)}{\theta_n - \theta_1} \right\},
\]

\( \eta = \frac{\hat{\zeta} + 1}{\hat{\zeta} - 1} \) with \( \hat{\zeta} = \kappa(\hat{H}_{\text{opt}}) = \frac{\hat{\theta}_n + \lambda_{\text{opt}}}{\hat{\theta}_1 + \lambda_{\text{opt}}} \),

and \( \epsilon^\beta(\eta) \) is as defined in (4.21).

5 Numerical examples

In this section, we will present several numerical examples by MATLAB (R2011b) to support our main theoretical conclusions in Theorem 4.2 and Remark 4.2. For that purpose, we choose medium size \( n \) but employ the sophisticated MATLAB routine, namely \texttt{trust} which is available in MATLAB 7.0 (R14) (but not other versions), for the original (1.1) as well as for the projected one (3.3). Moreover, in order to control the numerical effects arising from the loss of orthogonality of \( Q_k \) from the Lanczos process, we use the \texttt{Lanczos with full reorthogonalization} [4, Algorithm 7.2] to generate \( Q_k \). We shall compare the errors
\[
f(\mathbf{s}_k) - f(\mathbf{s}_{\text{opt}}) \quad \text{and} \quad \| \mathbf{s}_k - \mathbf{s}_{\text{opt}} \|_2
\]
with their corresponding bounds given in Theorem 4.2.

In constructing the testing matrices \( H \), without loss of generality we simply take \( H \) to be diagonal with translated Chebyshev zero nodes (to be defined) on the diagonal. We note that the \( n \) zero nodes and the \( n+1 \) extreme nodes of the \( n \)th Chebyshev polynomial \( \mathcal{T}_n \) in \([−1,1]\) are given respectively by
\[
t_{jn} = \cos \left( \frac{(2j-1)\pi}{2n} \right) \quad (1 \leq j \leq n) \quad \text{and} \quad \tau_{jn} = \cos \left( \frac{j\pi}{n} \right) \quad (0 \leq j \leq n).
\]

Given an interval \([a,b]\), the linear transformation
\[
z = \varpi(t - \xi) \quad (5.1)
\]
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 \( n \)th 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,
\[
t_{jn}^{\text{tr}} = \varpi(t_{jn} - \xi) \quad (1 \leq j \leq n) \quad \text{and} \quad \tau_{jn}^{\text{tr}} = \varpi(\tau_{jn} - \xi) \quad (0 \leq j \leq n).
\]
The reason behind choosing such matrices \( H \) is that the resulting linear systems \( H_{opt} s = -g \) are the hardest for CG, MINRES, and GMRES for a fixed \( \kappa(H_{opt}) \) as confirmed by the theoretical analysis in \([10, 11]\). In summary, we take

\[
H = \text{diag}(t_{1n}, \ldots, t_{nn}), \quad \text{or} \quad H = \text{diag}(\tau_{0n-1}, \ldots, \tau_{n-1n-1}),
\]

(5.2a)

\[
g = e := [1, \ldots, 1]^T.
\]

(5.2b)

In the examples that follow, we will examine the observed \( f(s_k) - f(s_{opt}) \) and \( \|s_k - s_{opt}\|_2 \) together with their upper bounds by Theorem 4.2 and also the upper bound approximations given by (4.33).

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 \([-10, 10]\).

**Example 5.1.** Setting \( \Delta = 1, n = 500 \) and \( [a, b] = [-10, 10] \), with (5.2a) and (5.2c) we get \( \lambda_{opt} \approx 25.3775 \) and \( \eta \approx 2.5378 \) by MATLAB’s trust, and with (5.2b) and (5.2c) we get \( \lambda_{opt} \approx 25.3826 \) and \( \eta \approx 2.5383 \) also by trust. Figure 5.1 plots the observed
The constraint \(\tau_k \leq \delta_k \) provide the sharpest estimates among the three types of sub-optimal polynomials,

2) both (4.23) and (4.16) are quite satisfactory, but the upper bounds in (4.17) delivered by the MINRES polynomials are the most overestimated, and

3) interestingly, the approximations given by (4.33) provide very accurate estimates for \(f(s_k) - f(s_{opt})\) and \(\|s_k - s_{opt}\|_2\). This shows that the estimates in Theorem 4.2 and Remark 4.2 are quite effective and practical.

Example 5.2. In this example, we continue the numerical evaluation of the bounds of Theorem 4.2, and evaluate the estimates of (4.25a) and (4.25b) but let \(\Delta\) vary. We still use the 500th Chebyshev zeros nodes (i.e., \([a, b] = [-1, 1]\)) as the diagonal entries of \(H\) and \(g = e\). In the left and right subfigures in Figure 5.2, we plot the observed \(f(s_k) - f(s_{opt})\) and \(\|s_k - s_{opt}\|_2\) together with the upper bounds in Theorem 4.2 and also the approximations given by (4.33), where

\[
\eta = \eta(\Delta) = \frac{1}{2} + \frac{\sqrt{\pi} \Delta}{\sqrt{k+1}(\eta)} - \frac{\ln(\frac{\sqrt{\pi} \Delta}{\sqrt{k+1}(\eta)})}{\eta - \theta_1},
\]

\[
\lambda_{opt}(\Delta) = \lambda_{opt}(\Delta) = \frac{1}{2} \left( \frac{2n-1}{2n} \right) \pi \approx 1 \quad \text{and} \quad \eta(\Delta) = 1 + \frac{\theta_1 \Delta^2 + \lambda_{opt}(\Delta)}{\theta_n - \theta_1} \Delta^2,
\]

where \(\lambda_{opt}(\Delta)\) is the Lagrangian multiplier associated with (5.3) and changes as \(\Delta\) varies. In this example, we have

\[
\theta_1 = \cos \left( \frac{(2n-1)\pi}{2n} \right) \approx -1 + \frac{\pi^2}{8n^2}, \quad \theta_n = \cos \left( \frac{\pi}{2n} \right) \approx 1 - \frac{\pi^2}{8n^2},
\]

and thus \(\eta(\Delta) \approx \frac{\lambda_{opt}(\Delta)}{\Delta^2}\). It is easy to see that if \(\frac{d\lambda_{opt}(\Delta)}{d\Delta} < \frac{2\lambda_{opt}(\Delta)}{\Delta^2}\), then \(\eta(\Delta)\) is a decreasing function of \(\Delta\), i.e., the bigger \(\Delta\) is, the smaller \(\eta(\Delta)\) becomes and thus the slower the convergence of \(s_k\) will be. This is well reflected by the slopes of the curves in Figure 5.2. In Figure 5.3, we plot \(\eta(\Delta)\) and \(\lambda_{opt}(\Delta)\) as the functions of \(\Delta\).
Example 5.3. We last 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 5.4 the observed $f(s_k) - f(s_{opt})$ and $\|s_k - s_{opt}\|_2$ together with their upper bounds according to Theorem 4.2 and also the approximations given by (4.33), with respect to $k$ for the two cases $b = 50$ and $b = 100$.

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

$$\zeta_k = \min \left\{ \frac{\sqrt{\pi} \Delta}{\mathcal{R}_{k+1}(\eta)}, \frac{\|g\|_2}{\mathcal{R}_{k+1}(\eta)}, \frac{2\|g\|_2 \epsilon_{\max}^2(\eta)}{\theta_n - \theta_1} \right\} = \frac{2\|g\|_2 \epsilon_{\max}^2(\eta)}{\theta_n - \theta_1}$$

for each computed Lanczos step $k$; in other words, (4.23) via the best polynomial for $\frac{1}{x-\eta}$ delivers the smallest values for $f(s_k) - f(s_{opt})$ and $\|s_k - s_{opt}\|_2$ among the three types of sub-optimal polynomials.

Also, a similar pattern as in Figure 5.2 is observed: $f(s_k) - f(s_{opt})$ and $\|s_k - s_{opt}\|_2$
and their upper bounds go to 0 faster as \( b \) get smaller. Indeed, in this example, we have \( \bar{\nu} = \frac{b}{2} \) and

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

and thus \( \eta \approx 1 + 2\frac{\lambda_{\text{opt}}}{b} \), which decreases as \( b \) increases. By checking the detailed numerical results from TLTRS, we find

\[
\lambda_{\text{opt}} \approx \begin{cases}
11.27, & b = 50, \\
8.75, & b = 100,
\end{cases}
\quad \text{and} \quad
\eta \approx \begin{cases}
1.46, & b = 50, \\
1.18, & b = 100.
\end{cases}
\]

![Figure 5.4: Example 5.3.](image)

Figure 5.4: Example 5.3. \( H \in \mathbb{R}^{500 \times 500} \) with translated Chebyshev zero nodes on two intervals \([a, b] = [0, 50]\) and \([a, b] = [0, 100]\).

6 Concluding remarks

In this paper, we have performed a convergence analysis for the truncated Lanczos approach (TLTRS) proposed in [6] for the standard trust-region subproblem. Mimicking the classical Rayleigh-Ritz procedure in the eigenvalue computations, TLTRS first projects a large-scale TRS into a much smaller TRS using the (preconditioned) Lanczos process, and then solves the smaller TRS by the Moré-Sorensen algorithm or some of its modifications. It is interesting to point out that, specially, when \( M = I \) and \( g = 0 \), TRS reduces to the standard symmetric eigenvalue problem which is often solved by the Lanczos method (e.g., [4, 18]). In that special case, the global optimal value \( f(s_{\text{opt}}) \) of (1.1) is the smallest eigenvalue \( \theta_1 \) of \( H \) while the global optimal value \( f(s_k) = \hat{f}(h_k) \) of the projected problem (3.3) is the smallest Ritz value \( \sigma_1 \). Elegant theoretical error bounds concerning the eigenvalues \( \theta_i \)/eigenvectors and the Ritz values \( \sigma_i \)/Ritz vectors from the Lanczos method have been established (e.g., [13, 18, 23]). The Chebyshev polynomials of the first kind have been playing an important role in these elegant theoretical results.
By making use of the special structure and optimality conditions of TRS, this paper addresses an important theoretical question of how good is the projected TRS in approximating the original problem: we have established error estimates for the approximate objective value $f(s_k)$ (cf. Ritz value) as well as the approximate solution (cf. Ritz vector) $s_k$ to the 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 estimates turn out to be quite sharp and can be effectively estimated using roughly $O(k^2)$ extra flops, and therefore they can be used to devise practical stopping criteria for TLTRS and GLTR. Finally numerical examples are presented to support our theoretical analysis.

References


