Finding a Global Optimal Solution for a Quadratically Constrained Fractional Quadratic Problem with Applications to the Regularized Total Least Squares

Amir Beck, Aharon Ben-Tal*
Faculty of Industrial Engineering and Management
Technion, Israel Institute of Technology, Haifa 3200, Israel

Marc Teboulle†
School of Mathematical Sciences
Tel-Aviv University, Ramat-Aviv 69978, Israel

July 6, 2005

Abstract

We consider the problem of minimizing a fractional quadratic problem involving the ratio of two indefinite quadratic functions, subject to a two sided quadratic form constraint. This formulation is motivated by the so-called Regularized Total Least Squares problem (RTLS). A key difficulty with this problem is its nonconvexity, and all current known methods to solve it are only guaranteed to converge to a point satisfying first order necessary optimality conditions. We prove that a global optimal solution to this problem can be found by solving a sequence of very simple convex minimization problems parameterized by a single parameter. As a result, we derive an efficient algorithm that produces an $\epsilon$-global optimal solution in a computational effort of $O(n^3 \log \epsilon^{-1})$. The algorithm is tested on problems arising from the inverse Laplace transform and image deblurring. Comparison to other well known RTLS solvers illustrate the attractiveness of our new method.

*The research was partially supported by BSF grant #2002038
†The research was partially supported by BSF grant #2002010
1 Introduction

In this paper we consider the problem of minimizing a fractional quadratic function subject to a quadratic constraint:

$$\min_{x \in F} \frac{f_1(x)}{f_2(x)},$$  \hspace{1cm} (1)

where

$$f_i(x) = x^T A_i x - 2b_i^T x + c_i, \quad i = 1, 2,$$  \hspace{1cm} (2)

$A_1, A_2 \in \mathbb{R}^{n \times n}$ are symmetric matrices, $b_1, b_2 \in \mathbb{R}^n$, $c_1, c_2 \in \mathbb{R}$ and $0 \leq L < U$. We do not assume that $A_1$ and $A_2$ are positive semidefinite and the only assumption required for the problem to be well defined is that $f_2(x)$ is bounded away from zero. We will discuss two cases of the feasible set $F$:

$$F_1 = \{x \in \mathbb{R}^n : L^2 \leq x^T Tx \leq U^2\},$$

where $T$ is a positive definite matrix and $U > L \geq 0$.

$$F_2 = \{x \in \mathbb{R}^n : x^T Bx \leq U^2\},$$

where $B$ is a positive semidefinite matrix and $U > 0$.

The major difficulty associated with problem (1) is the nonconvexity of the objective function, and in the case of $F_1$, also the nonconvexity of the feasible set.

The main motivation for considering problem (1) comes from the so called Regularized Total Least Squares (RTLS) problem. Many problems in data fitting and estimation give rise to an overdetermined system of linear equations $Ax \approx b$, where both the matrix $A \in \mathbb{R}^{m \times n}$ and the vector $b \in \mathbb{R}^m$ are contaminated by noise. The Total Least Squares (TLS) approach to this problem [9, 10, 15] is to seek a perturbation matrix $E \in \mathbb{R}^{m \times n}$ and a perturbation vector $r \in \mathbb{R}^m$ that minimize $\|E\|^2 + \|r\|^2$ subject to the consistency equation $(A + E)x = b + r$ (here and elsewhere in this paper a matrix norm is always the Frobenius norm and a vector norm is the Euclidean one). The TLS approach was extensively used in a variety of scientific disciplines such as signal processing, automatic control, statistics, physics, economics, biology and medicine (see e.g., [15] and references therein). The TLS problem has essentially an explicit solution, expressed by the singular value decomposition (SVD) of the augmented matrix $(A, b)$.

Regularization of the TLS solution is required in the case where $A$ is nearly rank deficient. Such problems arise, for example, from the discretization of ill-posed problems such as integral equations of the first kind (see e.g., [8, 13] and references therein). In these problems the TLS solution can be physically meaningless and thus regularization is employed in order to stabilize the solution.

Regularization of the TLS solution was addressed by several approaches: truncation methods [5, 13], Tikhonov regularization [8] and recently by introducing a quadratic constraint [20, 11, 8]. All the above methods are still trapped in the nonconvexity of the problem.
and thus are not guaranteed to converge to a global optimum. At best, they are proven to converge to a point satisfying first order necessary optimality condition. In contrast, in this paper, we develop an efficient algorithm which finds the global optimal solution by converting the original problem into a sequence of very simple convex optimization problems parameterized by a single parameter $\alpha$. The optimal solution corresponds to a particular value of $\alpha$, which can be found by a simple one dimensional search. The algorithm finds an $\epsilon$-optimal solution $x^*$ of (1), i.e.,

$$\frac{f_1(x^*)}{f_2(x^*)} \leq \min_{x \in \mathcal{F}} \frac{f_1(x)}{f_2(x)} + \epsilon,$$

in a computational effort of order $O(n^3 \log \left(\frac{1}{\epsilon}\right))$.

The paper is organized as follows. In the next section, we show how to recover the formulation of the RTLS problem as a quadratically constrained fractional quadratic problem. Section 3 describes a schematic algorithm designed to solve (1) for general quadratic functions $f_1$ and $f_2$ that provides the starting point of the analysis and the main results that are developed in Section 4. In Section 5 we return to the RTLS problem and give a detailed algorithm (RTLSC) for its solution. In order to illustrate the performance of algorithm RTLSC, two problems from the ”Regularization Tools” [13] are employed: a problem that arises from the discretization of the inverse Laplace transform and an image deblurring problem. These numerical examples are reported in Section 6 where we also compare the performance of our algorithm RTLSC with other well known RTLS solvers. Some useful technical results used throughout the paper are collected in the appendix.

## 2 The RTLS Problem

In this section we show how to recover a known formulation of the RTLS problem as a quadratically constrained fractional quadratic programming. This result is well known [9, 15, 20]. However, we believe that the derivation we give below is simpler. The RTLS problem as stated in [20] is

$$\min_{E, r, x} \|E\|^2 + \|r\|^2$$

s.t. \((A + E)x = b + r\)

$$x \in \mathcal{F}_2$$

(3)

To show that the RTLS problem (3) is a special case of problem (1), let us write (3) as

$$\min_{x \in \mathcal{F}_2} \min_{E, r: (A + E)x = b + r} \|E\|^2 + \|r\|^2$$

(4)

Next, fix $x \in \mathcal{F}_2$ and consider the inner minimization problem in (4). Denote $w = vec(E, r)$ where, for a matrix $M$, $vec(M)$ denotes the vector obtained by stacking the columns of $M$. The linear constraint (in $E$ and $r$) $(A + E)x = b + r$ can be written as $Q_x w = b - Ax$ where
\[ Q_x = \begin{pmatrix} \tilde{x}^T & 0 & \ldots & 0 \\ 0 & \tilde{x}^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{x}^T \end{pmatrix} \]

and \( \tilde{x} = (x^T, -1)^T \). Thus, the inner minimization problem in (4) takes the form

\[ \min_{Q_x w = b - Ax} \|w\|^2. \] (5)

Using the KKT conditions, it is easy to see that the solution of (5) is attained at \( w = Q_x (Q_x Q_x^T)^{-1} (b - Ax) \) and as a result the optimal value of problem (5) is equal to

\[ (b - Ax)^T (Q_x Q_x^T)^{-1} (b - Ax). \]

Since \( Q_x Q_x^T = \|\tilde{x}\|^2 I \) we deduce that the value of the inner minimization problem (5) is equal to \( \frac{\|Ax - b\|^2}{\|x\|^2} = \frac{\|Ax - b\|^2}{\|x\|^2 + 1} \). Consequently, the value of the RTLS problem (3) reduces to

\[ \min_{x \in \mathcal{F}_2} \frac{\|Ax - b\|^2}{\|x\|^2 + 1}, \] (6)

which is indeed a special case of problem (1).

### 3 A Schematic Algorithm

We consider problem (1) and in the sequel assume

**Assumption 1** \( f_2 \) is bounded below on \( \mathcal{F} \) by a positive number \( N \).

Let \( m \) and \( M \) be numbers such that

\[ m \leq \min_{x \in \mathcal{F}} \frac{f_1(x)}{f_2(x)} \leq M. \] (7)

Such bounds are easy to find, see Subsection 4.3.

**Remark 3.1** For the RTLS problem (6), assumption 1 is trivially satisfied for \( N = 1 \). The lower bound \( m \) can be chosen as 0 and \( M \) can be taken to be \( f(0) = \|b\|^2 \). Although both denominator and nominator in the RTLS problem (6) are convex, this property does not make the problem simpler since the quotient of convex functions is not necessarily convex.

A simple observation, that goes back to Dinkelbach [4], and that will enable us to solve (1) is the following,

**Observation:** The following two statements are equivalent

1. \( \min_{x \in \mathcal{F}} \frac{f_1(x)}{f_2(x)} \leq \alpha. \)

2. \( \min_{x \in \mathcal{F}} \{f_1(x) - \alpha f_2(x)\} \leq 0. \)
Using the above observation, we can solve (1) by the following schematic bisection algorithm.

**Schematic Algorithm**

**Initial Step:** Set \( lb_0 = m \) and \( ub_0 = M \).

**General Step:** For every \( k \geq 1 \):

1. Define \( \alpha_k = \frac{lb_{k-1} + ub_{k-1}}{2} \).

2. Calculate \( \beta_k = \min_{x \in F} \{ f_1(x) - \alpha_k f_2(x) \} \).
   
   (a) If \( \beta_k \leq 0 \) then define \( lb_k = lb_{k-1} \) and \( ub_k = \alpha_k \).

   (b) If \( \beta_k > 0 \) then define \( lb_k = \alpha_k \) and \( ub_k = ub_{k-1} \).

**Stopping Rule:** Stop at the first iteration \( k^* \) that satisfies \( ub_k^* - lb_k^* \leq \epsilon \).

**Output:**

\[
\begin{align*}
  x^* &\in \arg\min_{x \in F} \{ f_1(x) - ub_k^* f_2(x) \} \\
\end{align*}
\]

**Proposition 3.1** The schematic algorithm ends after \( \lceil \ln \left( \frac{M-m}{\epsilon} \right) / \ln(2) \rceil \) iterations with an output \( x^* \) that is an \( \epsilon \)-optimal solution of problem (1). More precisely,

\[
x^* \in F, \quad \alpha^* \leq \frac{f_1(x^*)}{f_2(x^*)} \leq \alpha^* + \epsilon,
\]

where \( \alpha^* = \min_{x \in F} \frac{f_1(x)}{f_2(x)} \).

**Proof:** The length of the initial interval is \( ub_0 - lb_0 = M - m \). By the definition of \( lb_k \) and \( ub_k \), we have that for every \( k \geq 1 \), \( ub_k - lb_k = \frac{1}{2}(ub_{k-1} - lb_{k-1}) \) and therefore \( ub_k - lb_k = (M - m) \left( \frac{1}{2} \right)^k \). From this it follows that \( k^* \), the number of iteration of the schematic algorithm, is the smallest integer \( k \) satisfying

\[
(M - m) \left( \frac{1}{2} \right)^k \leq \epsilon,
\]

which is equivalent to \( k \geq \lceil \ln \left( \frac{M-m}{\epsilon} \right) / \ln(2) \rceil \). By (8), \( x^* \) is feasible, i.e., \( x^* \in F \). Also, by the definition of the bisection process we have that \( lb_k \leq \alpha^* \leq ub_k \). By (8) we have that \( lb_k \leq \alpha^* \leq \frac{f_1(x^*)}{f_2(x^*)} \leq ub_k \) for every \( k \) and finally, since \( ub_k \leq lb_k + \epsilon \), the result follows. \( \square \)
Remark 3.2 By writing "min" and not "inf" in statements 1 and 2 of the observation and in the above scheme, we implicitly assumed that the minimum of the corresponding problems is attained (which is certainly the case when $\mathcal{F} = \mathcal{F}_1$). Otherwise, the inequalities in the statements of the observation should be replaced by strict inequalities and the schematic algorithm revised accordingly. The schematic algorithm then terminates with a point $x^*$, at which the objective value is at most $\epsilon$ away from the infimum. Thus henceforth we will assume that the minimum is attained.

To convert the schematic algorithm to a practical scheme we still need to address the following two questions:

1. How to choose the lower and upper bound $m$ and $M$?

2. How to solve the subproblem

$$\min_{x \in \mathcal{F}} \{f_1(x) - \alpha f_2(x)\}. \quad (9)$$

The first question is rather easy (see Subsection 4.3). The second one, is seemingly more difficult since problem (9), like the original problem (1), is nonconvex. In the next section we give complete answers to these two questions.

4 Analysis and Main Results

In Subsections 4.1 and 4.2 we show how to efficiently solve the subproblem (9). We first transform the problem (9) into a convex optimization problem by using the methodology of Ben-Tal and Teboulle [2]. We then show that the solution of the derived convex optimization problem consists of one eigenvector decomposition and solutions of at most two one-dimensional secular equations [17]. Finally, in Subsection (4.3) we show how to find the lower and upper bounds $m$ and $M$.

4.1 Solving the Subproblem in the Case $\mathcal{F} = \mathcal{F}_1$

In this subsection we consider the case in which the feasible set is equal to $\{x : L^2 \leq x^T Tx \leq U^2\}$, where $T$ is a positive definite matrix. Notice that in this case, the feasible set is compact and thus the minimum is always attained both in the original problem (1) and in the subproblem (9). First, we convert problem (9) to one with an Euclidean norm constraint by making the change of variables $s = T^{1/2}x$. The result is the following optimization problem:

$$\min_{L^2 \leq \|s\|^2 \leq U^2} \{f_1(T^{-1/2}s) - \alpha f_2(T^{-1/2}s)\}. \quad (10)$$

Using the following notation

$$\begin{align*}
\tilde{A} &= T^{-1/2}(A_1 - \alpha A_2)T^{-1/2}, \\
\tilde{b} &= T^{-1/2}(b_1 - \alpha b_2), \\
\tilde{c} &= c_1 - \alpha c_2,
\end{align*}$$

6
we obtain that problem (10) is the same as
\[
(P) : \min_{L \leq \|s\| \leq U} \{s^T \tilde{A}s - 2\tilde{b}^T s + \tilde{c}\}. \tag{11}
\]
\(\tilde{A}\) is symmetric and hence can be diagonalized by an orthogonal matrix \(U\), so that
\[
U^T \tilde{A} U = D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{12}
\]
where \(\lambda_1 \geq \lambda_1 \geq \ldots \geq \lambda_n\). Making the change of variables \(s = Uz\) we obtain that (11) is equivalent to
\[
\min_{L^2 \leq \|z\|^2 \leq U^2} \left\{ \sum_{j=1}^{n} (\lambda_j z_j^2 - 2f_j z_j) + \tilde{c} \right\}, \tag{13}
\]
where \(f = U^T b\). The following lemma will enable us to transform problem (13) into a convex optimization problem.

Lemma 4.1 Let \((z_1^*, z_2^*, \ldots, z_n^*)\) be an optimal solution of
\[
\min_{L^2 \leq \|z\|^2 \leq U^2} q(z),
\]
where
\[
q(z) = \sum_{j=1}^{n} (\lambda_j z_j^2 - 2f_j z_j). \tag{14}
\]
Then \(z_j^* f_j \geq 0\) for every \(j = 1, 2, \ldots, n\) for which \(f_j \neq 0\).

Proof: Since \(w = (z_1^*, z_2^*, \ldots, z_n^*)\) is optimal then it is in particular feasible, i.e., \(L^2 \leq \|w\|^2 \leq U^2\). An immediate result is that \((z_1^*, z_2^*, \ldots, z_{k-1}^*, -z_k^*, z_{k+1}^*, \ldots, z_n^*)\) is also feasible for every \(k = 1, 2, \ldots, n\). Since \(w\) is optimal we have that for every \(k = 1, 2, \ldots, n\):
\[
q(z_1^*, \ldots, z_n^*) \leq q(z_1^*, \ldots, z_{k-1}^*, -z_k^*, z_{k+1}^*, \ldots, z_n^*). \tag{15}
\]
Substituting (14) into (15) yields
\[
\sum_{j=1}^{n} (\lambda_j (z_j^*)^2 - 2f_j z_j^*) \leq \sum_{j=1, j \neq k}^{n} (\lambda_j (z_j^*)^2 - 2f_j z_j^*) + \lambda_k (-z_k^*)^2 + 2f_k z_k^*.
\]
Therefore, \(f_k z_k^* \geq 0\), and the result follows. \(\Box\)

Note that if \(f_j = 0\) for some \(j\), then the objective function \(q(z)\) is symmetric with respect to \(z_j\) and as a result we can arbitrarily restrict \(z_j\) to be nonnegative or nonpositive. In view of this and Lemma 4.1, we can make the change of variables
\[
z_j = \text{sign}(f_j) \sqrt{v_j}, \quad j = 1, 2, \ldots, n, \tag{16}
\]
where \(v_j \geq 0\). Substituting (16) into (13), we conclude that problem (9) is equivalent to the convex optimization problem
\[
\min_{v_j \geq 0} \left\{ \sum_{j=1}^{n} (\lambda_j v_j - 2|f_j| \sqrt{v_j}) + \tilde{c} : L^2 \leq \sum_{j=1}^{n} v_j \leq U^2 \right\}. \tag{17}
\]
Proposition 4.1 Let \( \tilde{A} \in \mathbb{R}^{n \times n} \) be a symmetric matrix, \( \tilde{b} \in \mathbb{R}^n, \tilde{c} \in \mathbb{R} \) and let the spectral decomposition of \( \tilde{A} \) be given by \( \tilde{A} = UDU^T \) where \( D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \) and \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \). Then the global solution to the optimization problem

\[
\min_{L^2 \leq \|s\|^2 \leq U^2} \left\{ s^T \tilde{A}s - 2\tilde{b}^T s + \tilde{c} \right\}
\]

is given by \( s = Uz \), where

\[
z_j = \text{sign}(f_j) \sqrt{v_j}, \quad j = 1, 2, \ldots, n
\]

and \( v \) is the solution of the convex optimization problem (17).

Proposition 4.1 shows that the main step in the schematic algorithm (step 2) consists of solving the linearly constrained convex optimization problem (17). This will be done by solving the dual problem, since, as we are about to show, the latter requires the solution of at most two single-variable convex problems.

To develop the dual problem of (17), we assign a nonnegative multiplier \( \xi \) to the linear inequality constraint \( -\sum_{j=1}^n v_j + L^2 \leq 0 \) and a nonpositive multiplier \( \eta \) to the linear inequality constraint \( -\sum_{j=1}^n v_j + U^2 \geq 0 \) and form the Lagrangian of (17):

\[
L(v, \eta, \xi) = \sum_{j=1}^n \left( \lambda_j v_j - 2|f_j| \sqrt{v_j} \right) - \eta \left( \sum_{j=1}^n v_j - U^2 \right) + \xi \left( -\sum_{j=1}^n v_j + L^2 \right) \tilde{c}
\]

\[
= \sum_{j=1}^n \left( (\lambda_j - \eta - \xi) v_j - 2|f_j| \sqrt{v_j} \right) + \eta U^2 + \xi L^2 + \tilde{c}. \tag{18}
\]

Differentiating (18) with respect to \( v_j \) and equating to zero, we obtain

\[
v_j = \frac{f_j}{(\lambda_j - \eta - \xi)^2}, \quad j = 1, 2, \ldots, n, \tag{19}
\]

subject to the conditions \( \eta + \xi \leq \lambda_n, \eta \leq 0 \) and \( \xi \geq 0 \). Thus, the dual objective function is given by

\[
\inf_{v_j \geq 0} L(v, \eta, \xi) = \begin{cases} h(\eta, \xi) & \text{if } \eta - \xi > -\lambda_n, \eta \leq 0, \xi \geq 0 \\ -\infty & \text{otherwise,} \end{cases}
\]

where

\[
h(\eta, \xi) = -\sum_{j=1}^n \frac{f_j^2}{\lambda_j - \eta - \xi} + \eta U^2 + \xi L^2 + \tilde{c}
\]

and the dual problem of (17) is

\[
(D) : \max_{\eta, \xi} \{ h(\eta, \xi) : \eta + \xi < \lambda_n, \eta \leq 0, \xi \geq 0 \}.
\]

From duality theory for convex optimization problems we have that [19, 3]

\[
\text{val}(P) = \text{val}(D),
\]

8
where \( \text{val}(P) \) (\( \text{val}(D) \)) denotes the optimal value of problem (P) (problem (D)). Now we note that the dual variables \( \eta \) and \( \xi \) cannot be both nonzero, since in that case we would have by the complementarity slackness condition that \( \sum_{j=1}^{n} v_j \) is equal to both \( U^2 \) and \( L^2 \), which is clearly a contradiction. As a result, instead of considering the problem (D) in two variables, we can consider the following two single-variable convex optimization problems (maximization of concave functions subject to a simple convex bound constraint):

\[
(D1) : \max_{\eta \leq \min(\lambda_n, 0)} \left( - \sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \eta} + \eta U^2 + \tilde{c} \right)_{h(\eta, 0)}
\]

and

\[
(D2) : \max_{0 \leq \xi < \lambda_n} \left( - \sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \xi} + \xi L^2 + \tilde{c} \right)_{h(0, \xi)}
\]

We thus obtain that in order to solve (D), we need to follow the following three steps:

1. Find a solution \( \eta \) of (D1).
2. Find a solution \( \xi \) of (D2).
3. If \( h(\eta, 0) > h(0, \xi) \) then the solution of (D) is \( (\eta, 0) \). Otherwise, the solution is \( (0, \xi) \).

Notice that both (D1) and (D2) are easy problems to solve since they consist of maximizing a concave function of a single variable. A very efficient algorithm for solving problems with an exact structure as (D1) and (D2) will be discussed at the end of this section.

We summarize our results on the solution of (11) in the following theorem.

**Theorem 4.1** Let \( \tilde{A} \in \mathbb{R}^{n \times n} \) be a symmetric matrix, \( \tilde{b} \in \mathbb{R}^n, \tilde{c} \in \mathbb{R} \) and let the spectral decomposition of \( \tilde{A} \) be given by \( \tilde{A} = U \Sigma U^T \) where \( \Sigma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \) with \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \). Then the solution to the optimization problem

\[
\min_{L^2 \leq \|s\|^2 \leq U^2} \left\{ s^T \tilde{A} s - 2 \tilde{b}^T s + \tilde{c} \right\}
\]

is \( s = Uz \), where \( z \in \mathbb{R}^n \) is given by

\[
z_j = \frac{f_j}{\lambda_j - \eta^* - \xi^*}, \quad j = 1, 2, \ldots, n,
\]

with \( (\eta^*, \xi^*) \) given by

\[
(\eta^*, \xi^*) = \begin{cases} 
(\bar{\eta}, 0) & \text{if } [\lambda_n > 0 \text{ and } h(\bar{\eta}, 0) > h(0, \bar{\xi})] \text{ or } \lambda_n \leq 0 \\
(0, \bar{\xi}) & \text{if } [\lambda_n > 0 \text{ and } h(\bar{\eta}, 0) \leq h(0, \bar{\xi})],
\end{cases}
\]

where \( \bar{\eta} \) and \( \bar{\xi} \) are the optimal solution of problems (D1) and (D2) respectively.
As was already mentioned, solving problems (D1) and (D2) is an easy task; to demonstrate this fact, let us consider the solution of (D1) in the case $\lambda_n \leq 0$ (all other instances can be similarly treated). In this case, (D1) takes the following form:

$$\max_{\eta < \lambda_n} \left\{ -\sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \eta} + \eta U^2 + \bar{c} \right\}.$$ 

Since $h_1(\eta) = h(\eta, 0)$ is continuous and strictly concave for $\eta < \lambda_n$ and also satisfies

$$\lim_{\eta \to -\infty} h_1(\eta) = -\infty, \quad \lim_{\eta \to \lambda_n^-} h_1(\eta) = -\infty,$$

we conclude that the maximum is obtained at a unique point $\eta < \lambda_n$ that satisfies $h_1'(\eta) = 0$.

Therefore, in this case we need to find the unique root of the following so-called secular equation [17]:

$$\eta < \lambda_n, \quad G(\eta) = U^2,$$

where

$$G(\eta) = \sum_{j=1}^{n} \frac{f_j^2}{(\eta - \lambda_j)^2}.$$  

Finding the unique root, which lies to the left of $\lambda_n$, of the secular equation (20) is a well studied problem (see e.g., [17, 7]). Specifically, Melman [17] transforms the problem into the following equivalent problem

$$G^{-1/2}(\eta) = U^{-1}$$  

for which Newton’s method exhibits global quadratic convergence. The algorithm is as follows.

**Algorithm SEC**

**Input:** $(f, \Lambda, U)$, where $f \in \mathbb{R}^n$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1 \geq \lambda_2 \ldots, \geq \lambda_n$ and $U > 0$.

**Output:** $\eta^* < \lambda_n$ that satisfies $|G(\eta^*) - U^2| < \epsilon_2$, where $G$ is defined in (21).

**Initial step:** $\eta_0 = \lambda_n - \epsilon_1$.

**General step:** for every $k \geq 0$,

$$\eta_{k+1} = \eta_k + \frac{G^{-1/2}(\eta_k) - U^{-1}}{G^{-3/2}(\eta_k)G''(\eta_k)}.$$

**Stopping rule:** Stop at the first iteration $k^*$ that satisfies $|G(\eta_{k^*}) - U^2| < \epsilon_2$. Set $\eta^* = \eta_{k^*}$.

In our implementation the tolerance parameters $\epsilon_1$ and $\epsilon_2$ take the values $\epsilon_1 = 10^{-4}$, $\epsilon_2 = 10^{-15}$. Melman’s algorithm solves the secular equation very fast (typically 5 or 6 iterations suffice to achieve 15 digits accuracy independently of $n$).

**Example:** To demonstrate the rate of convergence of algorithm SEC we consider problem (20) with $n = 100, \lambda_i = i, f_i = 1 (i = 1, 2, \ldots, 100)$ and $U = 1$. We compare algorithm SEC
Table 1: Quadratic rate of convergence of Melman’s Algorithm

with a simple bisection algorithm with initial interval $[-100, \lambda_n]$ and an identical stopping criteria as the one of algorithm SEC.

From Table 1 it is clear that the algorithm exhibits quadratic rate of convergence right from the very first iteration. The bisection algorithm terminated in this example after 55 iterations.

The dominant computational effort when solving the subproblem in the case $F = F_1$ are (i) the calculation of the matrices $T^{1/2}, T^{-1/2}$. (ii) The spectral decomposition of the matrix $A$. Each requires a computational effort of $O(n^3)$. By Proposition 3.1, the schematic algorithm requires to solve $O(\log \epsilon^{-1})$ subproblems in order to generate a $\epsilon$-global optimal solution. We thus conclude that the overall computational effort of the schematic algorithm is $O(n^3 \log \epsilon^{-1})$.

4.2 Solving the Subproblem in the Case $F = F_2$

Here we consider problem (9) in the case where the feasible set is $F_2 = \{x : x^TBx \leq U^2\}$, where $B$ is positive semidefinite but not positive definite. Thus, the subproblem in step 2 of the schematic algorithm under consideration here is

$$\beta^* = \min_{x^TBx \leq U^2} \{x^TAx - 2b^Tx + c\}, \quad (23)$$

where

$$A = A_1 - \alpha A_2, \quad b = b_1 - \alpha b_2, \quad c = c_1 - \alpha c_2.$$  

Notice that since $B$ is singular the feasible set $F_2$ is not compact and therefore the solution of the subproblem (23) might be $-\infty$. This issue is addressed in the following.

Lemma 4.2 Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, $U > 0$ and $B \in \mathbb{R}^{n \times n}$ be a positive semidefinite matrix. Then

1. If there exists $\lambda \geq 0$ such that $A + \lambda B \succ 0$, then the minimum of (23) is finite: $\beta^* > -\infty$. 

11
2. If no \( \lambda \geq 0 \) exists such that \( A + \lambda B \succeq 0 \), then the minimum of (23) is not finite.

**Proof:** The optimal value \( \beta^* \) of problem (23) is finite if and only if the following statement is true:

\[
\exists \mu \in \mathbb{R}, \quad x^T B x \leq U^2 \Rightarrow x^T A x + 2b^T x + c \geq \mu,
\]

which by the S-lemma (see Lemma A.1 in the appendix), is equivalent to

\[
\exists \mu \in \mathbb{R}, \lambda \in \mathbb{R}^+, \quad \begin{pmatrix} A & -b \\ -b^T & c - \mu \end{pmatrix} \succeq \lambda \begin{pmatrix} -B & 0 \\ 0 & U^2 \end{pmatrix},
\]

and which can also be written as

\[
\exists \mu \in \mathbb{R}, \lambda \in \mathbb{R}^+, \quad \begin{pmatrix} A + \lambda B & -b \\ -b^T & c - \mu - U^2 \end{pmatrix} \succeq 0.
\]

(25)

Since a necessary condition for the validity of (25) is that there exists a \( \lambda \geq 0 \) such that \( A + \lambda B \succeq 0 \), we conclude that the second statement of the lemma is proven. Moreover, if there exists a \( \lambda_0 \geq 0 \) such that \( A + \lambda_0 B \succ 0 \) then taking \( \mu_0 < c - U^2 - b^T(A + \lambda_0 B)^{-1}b \)
we have by Schur’s complement (Lemma A.2) that the LMI (25) is satisfied for \( \lambda = \lambda_0 \) and \( \mu = \mu_0 \) and therefore \( \beta^* > -\infty \) and the first statement of the lemma is proven. \( \Box \)

Notice that the only case not covered by Lemma 4.2 is the case where there is a \( \lambda \geq 0 \) such that \( A + \lambda B \succeq 0 \) but there does not exists a \( \lambda \geq 0 \) such that \( A + \lambda B \succ 0 \). Later on, we will see that we can ignore this case.

In the next result we find equivalent conditions for the finiteness of the minimization problem (23) that can be easily checked and analyzed.

**Lemma 4.3** Let \( A \in \mathbb{R}^{n \times n} \) be a symmetric matrix and let \( B \in \mathbb{R}^{n \times n} \) be a positive semidefinite matrix of rank \( r \). Denote by \( F \) the \( n \times (n - r) \) matrix whose columns are a basis for the null space of \( B \). Then the following two statements are equivalent

1. There exists \( \lambda \geq 0 \) such that \( A + \lambda B \succ 0 \).

2. \( F^T A F \succ 0 \).

**Proof:** First, since \( B \succeq 0 \), statement 1 is equivalent to the same statement without the sign constraint on \( \lambda \):

\[
\exists \lambda \in \mathbb{R}, \quad A + \lambda B \succeq 0.
\]

By Finsler’s Theorem (see Theorem A.1 in the appendix), this condition is equivalent to the following statement

\[
x^T A x > 0, \quad \text{for every } x \neq 0 \text{ such that } x^T B x = 0.
\]

(26)

Now, since \( B \succeq 0 \), we have that \( x^T B x = 0 \) is equivalent to \( x \in \text{Null}(B) \). Thus, (26) is equivalent to

\[
x^T A x > 0, \quad \text{for every } x \neq 0 \text{ such that } x \in \text{Null}(B),
\]

12
which is equivalent to saying that $F^T A F \succ 0$. □

A direct consequence of lemmata 4.3 and 4.2 is that if $F^T A F \succ 0$ \[(27)\]
then $\beta^* > -\infty$ and if $F^T A F$ is not positive semidefinite (i.e., has at least one negative eigenvector) then $\beta^* = -\infty$. In the case where condition (27) is satisfied we can simultaneously diagonalize $A$ and $B$ (see Appendix B) and therefore we can continue with the hidden convexity argument.

Let $C$ be a nonsingular matrix that simultaneously diagonalizes $A$ and $B$:

\[
C^T B C = \text{diag}(1, 1, \ldots, 1, 0, 0, \ldots, 0), \quad \text{r times } n - r \text{ times}
\]
\[
C^T A C = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r, 1, 1, \ldots, 1), \quad \text{n - r times}
\]

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ (see Appendix B for details). Making the change of variables $x = Cz$ we obtain that (23) is equivalent to

\[
\min \left\{ \sum_{j=1}^{r} \lambda_j z_j^2 + \sum_{j=r+1}^{n} z_j^2 - 2 \sum_{j=1}^{n} f_j z_j + c : \sum_{j=1}^{r} z_j^2 \leq U^2 \right\},
\]

where $f = C^T b$. The same argument as in Lemma 4.1 shows that we can make the change of variables

\[ z_j = \text{sign}(f_j) \sqrt{\nu_j}, \quad j = 1, 2, \ldots, n, \]

where $\nu_j \geq 0$. We obtain the following equivalent convex optimization problem

\[
\min_{\nu_j \geq 0} \left\{ \sum_{j=1}^{r} \left( \lambda_j v_j - 2|f_j|\sqrt{\nu_j} \right) + \sum_{j=r+1}^{n} \left( v_j - 2|f_j|\sqrt{\nu_j} \right) + c : \sum_{j=1}^{r} v_j \leq U^2 \right\}.
\]

To develop the dual problem of (29), we assign a nonpositive multiplier $\lambda$ to the linear inequality constraint $-\sum_{j=1}^{r} v_j + U^2 \geq 0$ and form the Lagrangian of (29) given by

\[
L(v, \eta, \xi) = \sum_{j=1}^{r} \left( \lambda_j v_j - 2|f_j|\sqrt{\nu_j} \right) + \sum_{j=r+1}^{n} \left( v_j - 2|f_j|\sqrt{\nu_j} \right) - \lambda \left( \sum_{j=1}^{r} v_j - U^2 \right) + c
\]
\[
= \sum_{j=1}^{r} \left( (\lambda_j - \lambda) v_j - 2|f_j|\sqrt{\nu_j} \right) + \sum_{j=r+1}^{n} \left( v_j - 2|f_j|\sqrt{\nu_j} \right) + \lambda U^2 + c.
\]

Differentiating (18) with respect to $v_j$ and equating to zero, we obtain

\[ v_j = \frac{f_j^2}{(\lambda_j - \lambda)^2}, \quad j = 1, 2, \ldots, r, \]
\[ v_j = f_j^2, \quad j = r + 1, \ldots, n \]
subject to the condition $\lambda \leq \min\{\lambda_n, 0\}$. Thus, the dual objective function is given by

$$ h(\lambda) = \inf_{v_j \geq 0} L(v, \eta, \xi) = \begin{cases} -\sum_{j=1}^r \frac{f_j^2}{\lambda_j - \lambda} + \lambda U^2 + d & \lambda < \min\{\lambda_r, 0\} \\ -\infty & \text{otherwise} \end{cases} $$

where $d = c - \sum_{j=r+1}^n f_j^2$. The dual problem of (29) is therefore

$$(D) : \max_{\lambda \leq \min\{\lambda_r, 0\}} h(\lambda).$$

From duality theory for convex optimization problems we have that [19, 3]

$$\text{val}(P) = \text{val}(D).$$

The solution of (D) involves the solution of a single secular equation of the form (20). We summarize the above discussion in Theorem 4.2 below.

**Theorem 4.2** Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $B \in \mathbb{R}^{n \times n}$ a positive semidefinite matrix of rank $r$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Moreover, suppose that $F^T A F \succ 0$, where $F$ is an $n \times (n-r)$ matrix whose columns are an orthogonal basis for the null space of $B$. Let $C$ be a nonsingular matrix for which the following is satisfied:

$$C^T BC = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}, \quad C^T AC = \begin{pmatrix} \Lambda & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. Then the solution to the optimization problem

$$\min_{x^T Bx \leq U^2} \{x^T A x - 2b^T x + c\}$$

is $x = Cz$, where $z \in \mathbb{R}^n$ is given by

$$z_j = \begin{cases} \frac{f_j}{\lambda_j - \lambda} & j = 1, 2, \ldots, r, \\ \frac{f_j}{f_j} & j = r + 1, \ldots, n \end{cases} \quad (f = C^T b)$$

and $\lambda$ is the solution to the following maximization problem

$$\max_{\lambda \leq \min\{\lambda_r, 0\}} \left\{ -\sum_{j=1}^r \frac{f_j^2}{\lambda_j - \lambda} + \lambda U^2 \right\},$$

whose solution consists of at most one root finding of a secular equation with one variable of the form (20).

We will impose an additional assumption on the quadratic function $f_2(x)$

**Assumption 2** $f_2$ is a strongly convex function (i.e., $A_2 \succ 0$).
Note that assumption 2 is readily satisfied by the RTLS problem (6). Recall that in the schematic algorithm $A = A_1 - \alpha A_2$, so (27) is equivalent to
\[
F^T A_1 F - \alpha F^T A_2 F \succ 0.
\] (31)
$F$ is full column rank and, by Assumption 2, we have that $A_2$ is positive definite and as a consequence $F^T A_2 F$ is also positive definite. Multiplying (31) from the right and left by $Q = (F^T A_2 F)^{-1/2}$, we obtain the following equivalent LMI
\[
Q (F^T A_1 F) Q - \alpha I \succ 0.
\]
The last LMI is equivalent to $\alpha < \lambda_{\min}(Q (F^T A_1 F) Q)$. We summarize this in the following proposition.

**Proposition 4.2** Let $\bar{\alpha} = \lambda_{\min}(Q (F^T A_1 F) Q)$, where $Q = (F^T A_2 F)^{-1/2}$. Then the minimum of (23) is finite if $\alpha < \bar{\alpha}$, and equal to $-\infty$ if $\alpha > \bar{\alpha}$.

$\bar{\alpha}$ is of course an upper bound for the minimal value of the original problem (1) and thus, in the schematic algorithm, we will always take an upper bound $M$ that is of most $\bar{\alpha}$. We therefore conclude that throughout the schematic algorithm, we need to consider only subproblems with finite minimum that satisfies (27).

A similar argument to the one given in the case $F = F_1$ shows that in the case $F = F_2$ as well, the algorithm produces an $\epsilon$-global optimal solution in a computational effort of $O(n^3 \log \epsilon^{-1})$.

### 4.3 Finding the Bounds

In this section we present some suggestions for the lower and upper bounds $m$ and $M$ of the schematic algorithm. In the special case of the original RTLS problem, simpler bounds are derived in Section 5.

#### 4.3.1 The Case $F = F_1$

In this case the constraint is given by $L^2 \leq x^T T x \leq U^2$. From this it follows that $\|x\|^2 \leq U^2 / \lambda_{\min}(T)$. We can therefore bound the objective function of problem (1) as follows
\[
\begin{vmatrix}
 f_1(x) \\
 f_2(x)
\end{vmatrix} = \begin{vmatrix}
 x^T A_1 x - 2b_2^T x + c_1 \\
 x^T A_2 x - 2b_2^T x + c_2
\end{vmatrix} \leq \frac{1}{N} \begin{vmatrix}
 x^T A_1 x - 2b_1^T x + c_1
\end{vmatrix} \\
\leq \frac{1}{N} (\|x^T A_1 x\| + \|2b_1^T x\| + |c_1|) \leq \frac{1}{N} \left( U^2 \frac{\lambda_{\max}(A_1)}{\lambda_{\min}(T)} + 2 \frac{\|b_1\| U}{\sqrt{\lambda_{\min}(T)}} + |c_1| \right).
\]

Thus, we can choose $m$ and $M$ to be
\[
M = \frac{1}{N} \left( U^2 \frac{\lambda_{\max}(A_1)}{\lambda_{\min}(T)} + 2 \frac{\|b_1\| U}{\sqrt{\lambda_{\min}(T)}} + |c_1| \right), \quad m = -M.
\]
The only element in the definition of $m$ and $M$, which is not given explicitly is the positive number $N$, defined in Assumption 1. For the RTLS problem, where $f_2(x) = \|x\|^2 + 1$, we can take $N$ to be equal to 1. Also, for other problems we can define $N$ to be the optimal value of the minimization problem $\min_{\|x\| \leq U} \{ x^T A_2 x - 2b_2^T x + c_2 \}$.
4.3.2 The Case $\mathcal{F} = \mathcal{F}_2$

In this case the constraint is given by $\mathbf{x}^T \mathbf{B} \mathbf{x} \leq U^2$, where $\mathbf{B}$ is a positive semidefinite matrix. We consider the case where both assumptions 1 and 2 hold true and that $f_2$ is bounded below in $\mathbb{R}^n$. The upper bound can be taken as $M = \bar{\alpha}$, where $\bar{\alpha}$ is given in Proposition 4.2.

To find a lower bound $m$, we first make the change of variables $\mathbf{z} = \mathbf{x} - A_2^{-1} \mathbf{b}_2$ resulting with the following form of the objective function

$$z^T A_1 z - 2e^T z + f = \frac{z^T A_2 z + d}{z^T A_2 z + d},$$

(32)

where $d = c_2 - b_2^T A_2^{-1} b_2 > 0$, $e = b_1 - A_1 A_2^{-1} b_2$ and $f = c_1 + b_2^T A_2^{-1} A_1 A_2^{-1} b_2 - 2b_1^T A_2^{-1} b_2$.

The unconstrained minimum of the last expression (32) is a lower bound on the optimal value and we can lower bound it using a relaxation technique.

$$\min_{\mathbf{z}} \frac{z^T A_1 z - 2e^T z + f}{z^T A_2 z + d} \Rightarrow \min_{w} \frac{w^T A_2^{-1/2} A_1 A_2^{-1/2} w - 2e^T A_2^{-1/2} w + f}{\|w\|^2 + d} = \min_{w, t = \sqrt{d}} \frac{w^T A_2^{-1/2} A_1 A_2^{-1/2} w - 2\sqrt{d} e^T A_2^{-1/2} w t + f t^2}{\|w\|^2 + t^2} \geq \lambda_{\min} \left( A_2^{-1/2} A_1 A_2^{-1/2}, \frac{1}{\sqrt{d}} e^T A_2^{-1/2}, \frac{f}{d} \right)$$

Thus, we can take $m = \lambda_{\min} \left( A_2^{-1/2} A_1 A_2^{-1/2}, \frac{1}{\sqrt{d}} e^T A_2^{-1/2}, \frac{f}{d} \right)$.

5 A Detailed Algorithm for the RTLS problem

In this section we use the results obtained so far to write in full details the schematic algorithm of Section 3 as applied to the RTLS problem:

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) \equiv \frac{\|A \mathbf{x} - \mathbf{b}\|^2}{\|\mathbf{x}\|^2 + 1} : \|\mathbf{L} \mathbf{x}\| \leq U \right\}.$$

(33)

We call this algorithm RTLSC.

The RTLSC algorithm solves at each iteration a subproblem of the form:

$$\min \left\{ \mathbf{x}^T Q \mathbf{x} - 2d^T \mathbf{x} : \|\mathbf{L} \mathbf{x}\| \leq U \right\}.$$

(34)

The detailed algorithm SUBP for solving the latter problem is explicitly written below. It invokes three procedures:

- SDG - an algorithm for simultaneous diagonalization of two matrices, one of which is positive definite (see Appendix B).
• SDGP - an algorithm for simultaneous diagonalization of two matrices, one of which is positive semidefinite (see Appendix B).

• SEC - Melman’s algorithm for solving secular equations given in Section 4.

Algorithm SUBP
Input: \((Q, d, L, U, F)\), where \(Q \in \mathbb{R}^{n \times n}\) is a symmetric matrix, \(d \in \mathbb{R}^n\), \(L \in \mathbb{R}^{r \times n}(r \leq n)\) is a full rank matrix, \(U > 0\) and \(F \in \mathbb{R}^{n \times (n-r)}\) is a matrix whose columns are an orthogonal basis for the null space of \(L\).

Output: \((x^*, \mu)\). \(x^*\) is an optimal solution to problem (34) and \(\mu\) is the corresponding optimal value.

1. If \(r < n\) then call algorithm SDG with input \((A, L^T L, F)\) and obtain an output \((C, \Lambda)\).
   Else call algorithm SDGP with input \((A, L^T L)\) and obtain an output \((C, \Lambda)\).

2. Set \(f = C^T d\).

3. If \(\lambda_r > 0\) and \(\sum_{j=1}^{r} \frac{f_j^2}{\lambda_j} < U^2\) then set \(\lambda^* = 0\). Else call algorithm SEC with input \((f, \Lambda, U)\) and obtain an output \(\lambda^*\).

4. Let \(v_j = \frac{f_j}{\lambda_j}, j = 1, \ldots, r\) and \(v_j = f_j, j = r + 1, \ldots, n\) \((\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r))\).

5. Set \(x^* = Cv\) and \(\mu = (x^*)^T Q x^* - 2f^T x^*\).

Algorithm RTLSC
Input: \((A, b, L, U, ub, \epsilon)\), where \(A \in \mathbb{R}^{m \times n}(m \geq n)\), \(b \in \mathbb{R}^m\), \(L \in \mathbb{R}^{r \times n}(r < n)\) has full row rank, \(U > 0\), \(ub > 0\) is an upper bound on the optimal function value and \(\epsilon > 0\) is a tolerance parameter.

Output: \(x^*\) - an \(\epsilon\) optimal solution of problem (33).

1. Set \(k = 0, lb_0 = 0, ub_0 = ub\).

2. Calculate a matrix \(F \in \mathbb{R}^{n \times (n-r)}\) whose columns are an orthogonal basis for the null space of \(L\).

3. While \(ub_k - lb_k > \epsilon\), do
   
   \(\alpha_k = \frac{lb_k + ub_k}{2}\).
   
   (b) Call algorithm SUBP with input \((A^T A - \alpha_k I, A^T b, L, U)\) and obtain an output \((x_k, \beta_k)\).
   
   (c) Calculate \(f_k = f(x_k)\).
   
   (d) If \(\beta_k + \|b\|^2 - \alpha_k > 0\) then
   
   \[lb_{k+1} = \alpha_k, ub_{k+1} = \min\{ub_k, f_k\},\]  \hspace{1cm} (35)
   
   else
   
   \[lb_{k+1} = lb_k, ub_{k+1} = \min\{\alpha_k, f_k\}.\]  \hspace{1cm} (36)
4. Define $x^* = x_m$, where $m$ is chosen so that $f_m = \min\{f_0, f_1, \ldots, f_{k-1}\}$.

Choice of lower and upper bounds: In the case where $L$ is square and nonsingular, the upper bound can be chosen as $ub = f(\tilde{x})$, where $\tilde{x}$ is any feasible point (such as $0$). A tight upper bound can be obtained by choosing $\tilde{x}$ as a solution of another method such as regularized least squares. In the rank deficient case, Proposition 4.2 implies that $\lambda_{\min}(F^TAF)$ is an upper bound on the optimal function value. Hence, an initial upper bound is given by $\min\{\lambda_{\min}(F^TAF), f(\tilde{x})\}$. This choice guarantees that all subproblems have a finite value.

Remark 5.1 Note that the update equations (35) and (36) for the upper bound $ub_k$ are different from the naive implementation suggested in the schematic algorithm of Section 3. The idea behind the revised update formulas is to incorporate the information gained at previous iterations in order to find better upper bounds. At each iteration we calculate a new feasible point $x_k$, which induces a new upper bound $f_k \equiv f(x_k)$ on the optimal function value. Thus, the update equation $ub_{k+1} = ub_k$ in the original schematic algorithm is converted to $ub_{k+1} = \min\{ub_k, f_k\}$. The following example demonstrates the advantage of using the new update equations.

Example: In this section we illustrate a single run of the RTLSC algorithm. We consider problem (33) with

$$ n = 2, A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, b = \begin{pmatrix} 10 \\ 25 \end{pmatrix}, L = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \rho = 10. $$

Table 2 describes the first six iterations of algorithm RTLSC. The initial upper bound $ub_0$ was chosen to be $f(0) = \|b\|^2 = 725$. Note that the decrease in the upper bound is very drastic at the first few iterations. The size of the interval $[lb_k, ub_k]$ decreases by a factor of 3000 between iteration 0 and iteration 1 (instead of a factor of 2 in the old update equations). The minimum value is equal to 0.047501 and it is reached after only three iterations. This run is typical in the sense that usually the algorithm converges to a point after very few iterations.

6 Numerical Examples

In order to test the performance of algorithm RTLSC, two problems from the ”Regularization Tools” [13] are employed: a problem that arises from the discretization of the inverse Laplace transform and an image deblurring problem. The following algorithms are tested:

- **RLS** - Regularized Least Squares. This is the solution to the problem

$$ \min\{\|Ax - b\|^2 : \|Lx\| \leq \rho\}, $$

implemented in the function l sqi from [13].
<table>
<thead>
<tr>
<th>k (# of iteration)</th>
<th>lb_k</th>
<th>ub_k</th>
<th>α_k</th>
<th>f_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>725</td>
<td>362.5</td>
<td>0.240383</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.243083</td>
<td>0.121541</td>
<td>0.047513</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.047513</td>
<td>0.023756</td>
<td>0.085005</td>
</tr>
<tr>
<td>3</td>
<td>0.023756</td>
<td>0.047513</td>
<td>0.035634</td>
<td>0.047501</td>
</tr>
<tr>
<td>4</td>
<td>0.035634</td>
<td>0.047501</td>
<td>0.041567</td>
<td>0.047501</td>
</tr>
<tr>
<td>5</td>
<td>0.041567</td>
<td>0.047501</td>
<td>0.044534</td>
<td>0.047501</td>
</tr>
<tr>
<td>6</td>
<td>0.044534</td>
<td>0.047501</td>
<td>0.046017</td>
<td>0.047501</td>
</tr>
</tbody>
</table>

Table 2: Single run of algorithm RTLSC

- **TTLS** - Truncated Total Least Squares originating from [5] and implemented in the function `ttls` from [13].
- **RTLSC** - Our algorithm from Section 5.
- **QEP** - Sima, Van Huffel and Golub’s solver for RTLS [20].

### 6.1 Inverse Laplace Transform

We consider the problem of estimating the function $f(t)$ from its given Laplace transform [21]:

$$
\int_0^\infty e^{-st}f(t)dt = \frac{2}{(s + 1/2)^3}.
$$

By means of Gauss-Laguerre quadrature, the problem reduces to a linear system $Ax = b$. This system and its solution $x_R$ are implemented in the function `ilaplace(n,3)` from [13]. The perturbed right-hand is generated by

$$
\tilde{b} = (A + \sigma E)x_R + \sigma e,
$$

where each component of $E$ and $e$ is generated from a standard normal distribution and $\sigma$ runs thorough the values 1e-1, 1e-2 and 1e-4. The matrix $L$ approximates the first-derivative operator implemented in the function `getl(n,1)` from [13]. Two cases are tested: $m = n = 20$ and $m = n = 100$. Table 3 describes the relative error $\|x - x_R\|/\|x_R\|$ averaged over 300 random realizations of $E$ and $e$.

The best results in each row are emphasized in boldface. The RTLSC and QEP methods give the best results in all but one case. The RLS also performed quite well. Note that the average relative error for the RTLSC and QEP solvers are equal. It is interesting to note that not only the average was the same but in fact for all 1800 simulations of QEP and RTLSC, the results were the same. Incidentally, this provides an experimental evidence to the
claim that QEP finds the global minimum, although such a theoretical claim was not proved in [20].

The CPU time in seconds of the three RTLS solvers averaged over 20 realizations of $E$ and $e$ is given in Table 4 below ($\sigma$ was fixed to be 1e-4). To make a fair comparison, we employed the same stopping rule for each of the methods: $\|x_{k+1} - x_k\|/\|x_k\| < 10^{-3}$.

<table>
<thead>
<tr>
<th>n</th>
<th>RTLSC</th>
<th>QEP</th>
<th>RLS</th>
<th>GR</th>
<th>TTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>5.6e-2</td>
<td>5.3e-2</td>
<td>2e-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.9e-1</td>
<td>2.1e-1</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.27</td>
<td>2.4</td>
<td>23.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>3.2</td>
<td>3.1</td>
<td>112.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>296</td>
<td>312</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: CPU time in seconds on a Pentium 4, 1.8Ghz.

It is clear from Table 4 that RTLSC and QEP are significantly faster than GR. Moreover, RTLSC and QEP require more or less the same running time.

### 6.2 Image Deblurring

We consider the problem of estimating a $32 \times 32$ twodimensional image obtained from the sum of three harmonic oscillations:

$$x(z_1, z_2) = \sum_{l=1}^{3} a_i \cos(w_{l,1}z_1 + w_{l,2}z_2 + \phi_l), \quad \left( w_{l,i} = \frac{2\pi k_{l,i}}{n} \right), \quad 1 \leq z_1, z_2 \leq 32$$

where $k_{l,i} \in \mathbb{Z}^2$ (see Fig.1, A). The specific values of the parameters are given in Table 5.

The image is blurred by atmospheric turbulence blur originating from [12] and implemented in the function `blur(n,3,1)` from [13].

The blurred image is generated by the relation (37) with $\sigma = 0.02$, which results with a highly noisy image (see Fig. 1, B).

**Choice of Regularization matrix:** We first ran algorithm RLS with standard regularization ($L = I$). The result is the poor image given in Fig. 1, C. We then chose $L$ as a
discrete approximation of the Laplace operator which is a two-dimensional convolution with the mask:

$$\begin{bmatrix}
-1 & -1 & -1 \\
-1 & 8 & -1 \\
-1 & -1 & -1 
\end{bmatrix}.$$  

This operator is standard in image processing [16]. With this choice of $L$, the RLS algorithm gave the much better image in Fig. 1, D. The above results demonstrate the importance of the choice of the regularization matrix. In the following experiments we use the nonstandard $L$. The result for algorithm TTLS is given in Fig. 1, E. A much improved result is obtained by our algorithm RTLSC (Fig. 1, F). Here again algorithm QEP gave the same result as algorithm RTLSC. Also, algorithm GR gave in this example the same image as RLS.

It is interesting to note that in this and many other examples algorithm RTLSC required only three iterations in order to produce quality reconstructions. As an illustration, Fig. 2 shows the result of the first three iterations of algorithm RTLSC. The function values of the images generated in iterations 1, 2 and 3 are 2.0934, 1.5715 and 1.5566 respectively. The difference between the first and second iteration is substantial. However, the image produced at the third iteration is almost identical to the image produced at the third iteration. Further iterations of RTLSC do not improve the image although the function value reduces to the minimal value 1.5234.
Figure 1: Results for different regularization solvers
Acknowledgments

We thank the associate editor and two anonymous referees for their constructive comments.
Appendix

A Known Results

Lemma A.1 (S-lemma [1]) Let $A$ and $B$ be $n \times n$ symmetric matrices and let $e, f \in \mathbb{R}^n$ and $g, h \in \mathbb{R}$. Assume that the quadratic inequality

$$x^T Ax + 2e^T x + g \geq 0$$

is strictly feasible i.e., there exists $\bar{x}$ such that $\bar{x}^T A \bar{x} + 2e^T \bar{x} + g > 0$. Then the quadratic inequality

$$x^T Bx + 2f^T x + h \geq 0$$

is a consequence of (38) if and only if there exists a nonnegative $\lambda$ such that

$$
\begin{pmatrix}
B & f \\
f^T & h
\end{pmatrix} \succeq \lambda
\begin{pmatrix}
A & e \\
e^T & g
\end{pmatrix}.
$$

Lemma A.2 (Schur’s complement [1]) Let

$$M = \begin{pmatrix}
A & B^T \\
B & C
\end{pmatrix}$$

be a symmetric matrix with $C \succ 0$. Then $M \succeq 0$ if and only if $\Delta_C \succeq 0$, where $\Delta_C$ is the Schur complement of $C$ in $M$ and is given by

$$\Delta_C = A - B^T C^{-1} B.$$

Theorem A.1 (Finsler’s Theorem [6]) Let $A$ and $B$ be symmetric $n \times n$ matrices. Then the quadratic inequality

$$x^T Bx > 0$$

is a consequence of the quadratic equality

$$x^T Ax = 0$$

if and only if there exists $\alpha \in \mathbb{R}$ such that $B - \alpha A \succ 0$.

B Algorithms for Simultaneous Diagonalization

In this Section we recover an algorithm for the simultaneous diagonalization of an $n \times n$ symmetric matrix $A$ and a positive semidefinite matrix $B \in \mathbb{R}^{n \times n}$ of rank $r(< n)$. We denote by $F$ the $n \times (n - r)$ matrix whose columns are an orthogonal basis for the null space of $B$ and assume that the condition

$$F^T A F \succ 0$$

(40)
is satisfied, which implies that the matrices \( A \) and \( B \) are simultaneously diagonalizable by a nonsingular matrix. This fact follows directly from [18, Theorem 6.2.2]. Here we explicitly recover the algorithm that follows from [18] for the special case where (40) is satisfied.

**Algorithm SDG**

**Input:** \((A, B, F)\), where \( A \in \mathbb{R}^{n \times n} \) is a symmetric matrix, \( B \in \mathbb{R}^{n \times n} \) is a positive semidefinite of rank \( r (r < n) \) and \( F \in \mathbb{R}^{n \times (n-r)} \) is a matrix whose columns are an orthogonal basis for the null space of \( B \).

**Condition:** \( F^T AF \succ 0 \).

**Output:** \((C, \Lambda)\). \( C \) is a nonsingular matrix and \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r) \) (\( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \)) is a diagonal matrix such that

\[
C^T BC = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}, \quad C^T AC = \begin{pmatrix} \Lambda & 0 \\ 0 & I_{n-r} \end{pmatrix}.
\]

1. Find a full row rank \( r \times n \) matrix \( L \) such that \( B = L^T L \).
2. Define \( M = L^T (LL^T)^{-1} \) (\( M \) is a right inverse of \( L \)). We have \( M^T BM = I_r \).
3. Define \( S = \begin{pmatrix} M - F(F^T AF)^{-1}F^T AM, & F \end{pmatrix} \). We have

\[
S^T BS = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}, \quad S^T AS = \begin{pmatrix} E & 0 \\ 0 & F^T AF \end{pmatrix},
\]

where \( E \) is an \( r \times r \) symmetric matrix.

4. Find an \( r \times r \) orthogonal matrix \( Q_1 \) such that \( Q_1^T EQ_1 = \Lambda \), where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r) \) with \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \).

5. Find an \( (n-r) \times (n-r) \) matrix \( Q_2 \) such that \( Q_2^T (F^T AF)Q_2 = I_{n-r} \) (this is possible since we assume that \( F^T AF \succ 0 \)).

6. Define

\[
C = S \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}.
\]

In the case where one of the matrices is positive definite, simultaneous diagonalization is always possible without any restrictions [14]. The procedure for simultaneous diagonalization in that case is much simpler and is given below

**Algorithm SDGP**

**Input:** \((A, B)\), where \( A \in \mathbb{R}^{n \times n} \) is a symmetric matrix and \( B \in \mathbb{R}^{n \times n} \) is a positive definite matrix.

**Output:** \((C, \Lambda)\). \( C \) is a nonsingular matrix and \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \) (\( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \)) is a diagonal matrix such that

\[
C^T BC = I, \quad C^T AC = \Lambda.
\]

\(^1\)This step can be done by e.g., Cholesky’s factorization. In some applications \( B \) is already given in that form.
1. Find a singular matrix $L$ such that $B = L^T L$. 

2. Calculate the spectral decomposition of $(L^T)^{-1} A L^{-1}$:

$$U^T((L^T)^{-1}A L^{-1})U = D,$$

where $U$ is an orthogonal matrix, $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

3. Set $C = L^{-1}U$, $\Lambda = D$.

References


