CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.2021.nla.01

Vol. **2**, No. 2, pp. 195-262
June 2021

Linear Constrained Rayleigh Quotient Optimization: Theory and Algorithms

Yunshen Zhou[1], Zhaojun Bai[2,*] and Ren-Cang Li[3]

[1] *Department of Mathematics, University of California, Davis, CA 95616, USA.*
[2] *Department of Computer Science and Department of Mathematics, University of California, Davis, CA 95616, USA.*
[3] *Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA.*

Received 16 July 2020; Accepted 31 December 2020

Abstract. We consider the following constrained Rayleigh quotient optimization problem (CRQopt):

$$\min_{v\in\mathbb{R}^n} v^{\mathrm{T}}Av \quad \text{subject to } v^{\mathrm{T}}v=1 \text{ and } C^{\mathrm{T}}v=b,$$

where A is an $n\times n$ real symmetric matrix and C is an $n\times m$ real matrix. Usually, $m\ll n$. The problem is also known as the constrained eigenvalue problem in literature since it becomes an eigenvalue problem if the linear constraint $C^{\mathrm{T}}v=b$ is removed. We start by transforming CRQopt into an equivalent optimization problem (LGopt) of minimizing the Lagrangian multiplier of CRQopt, and then into another equivalent problem (QEPmin) of finding the smallest eigenvalue of a quadratic eigenvalue problem. Although these equivalences have been discussed in literature, it appears to be the first time that they are rigorously justified in this paper. In the second part, we present numerical algorithms for solving LGopt and QEPmin based on Krylov subspace projection. The basic idea is to first project LGopt and QEPmin onto Krylov subspaces to yield problems of the same types but of much smaller sizes, and then solve the reduced problems by direct methods, which is either a secular equation solver (in the case of LGopt) or an eigensolver (in the case of QEPmin). We provide convergence analysis for the proposed algorithms and present error bounds. The sharpness of the error bounds is demonstrated by examples, although in applications the algorithms often converge much faster than the bounds suggest. Finally, we apply the new algorithms to semi-supervised learning in the context of constrained clustering.

AMS subject classifications: 65F15, 65F50, 90C20

Key words: Quadratic programming, quadratic eigenvalue problems, Lanczos algorithms, constrained clustering.

*Corresponding author. *Email addresses:* `yshzhou@ucdavis.edu` (Y. Zhou), `zbai@ucdavis.edu` (Z. Bai), `rcli@uta.edu` (R.-C. Li)

©2021 Global-Science Press

1 Introduction

In this paper, we are concerned with the following *linear constrained Rayleigh quotient* (CRQ) optimization:

$$\text{CRQopt:}\quad \begin{cases} \min\ v^{\mathrm{T}}Av, & (1.1\text{a})\\ \text{s.t. } v^{\mathrm{T}}v=1, & (1.1\text{b})\\ \qquad C^{\mathrm{T}}v=b, & (1.1\text{c}) \end{cases}$$

where $A\in\mathbb{R}^{n\times n}$ is symmetric, $C\in\mathbb{R}^{n\times m}$ has full column rank, and $b\in\mathbb{R}^{m}$. Necessarily $m<n$ but often $m\ll n$. We are particularly interested in the case where A is large and sparse and $b\neq 0$.

CRQopt (1.1) is also known as *the constrained eigenvalue problem*, a term coined in 1989 [10]. However, it had appeared in literature much earlier than that [15]. In that sense, CRQopt is a classical problem. However, past studies are fragmented with some claims, although often true, not rigorously justified or needed conditions to hold. In this paper, our goal is to provide a thorough investigation into this classical problem, including rigorous justifications of statements previously taken for granted in literature and addressing the theoretical subtleties that were not paid attention to. We also present a quantitative convergence analysis for the Krylov type subspace projection method, which we will also call the Lanczos algorithm, for solving large scale CRQopt (1.1).

1.1 Related works

CRQopt (1.1) has found a wide range of applications, such as ridge regression [5, 12], trust-region subproblem [27, 33], constrained least square problem [9], spectral image segmentation [6, 36], transductive learning [19], and community detection [28].

The first systematic study of CRQopt (1.1) belongs to Gander, Golub and von Matt [10]. Using the full QR and eigen-decompositions, they reformulated CRQopt (1.1) as an optimization problem of finding the minimal Lagrangian multiplier via solving a secular equation (in a way that is different from our secular equation solver in Appendix A). Alternatively, they also turned CRQopt (1.1) into an optimization problem of finding the smallest real eigenvalue of a quadratic eigenvalue problem (QEP). However, the equivalence between the QEP optimization and the Lagrangian multiplier problem was not rigorously justified in [10].

Numerical algorithms proposed in [10] are not suitable for large scale CRQopt (1.1) because they require a full eigen-decomposition of A. Later in [14], Golub, Zhang and Zha considered large and sparse CRQopt (1.1) but only with the homogeneous constraint, i.e, $b=0$. In this special case, CRQopt (1.1) is equivalent to computing the smallest eigenvalue of A restricted to the null space of C^{T}. An inner-outer iterative Lanczos method was proposed to solve the homogeneous CRQopt (1.1). In [41], Xu, Li and Schuurmans proposed a projected power method for solving CRQopt (1.1). The projected power method is an iterative method only involving matrix-vector products, and thus it is suitable for large and sparse CRQopt (1.1). However, its convergence is linear at best and often too slow. In [6], Eriksson, Olsson and Kahl reformulated CRQopt (1.1) into an eigenvalue optimization problem (see Appendix B for details). An algorithm based on the line search was used to find the optimal solution. This algorithm is suitable for CRQopt (1.1) with a large and sparse matrix A, but it is too costly because the smallest eigenvalue has to be computed multiple times during each line search action.

1.2 Contributions

Our study on CRQopt (1.1) begins with the standard approach of Lagrangian multipliers, as was taken in [10], to lead to an optimization problem of minimizing the Lagrangian multiplier of CRQopt, called LGopt (Section 2.2). Then LGopt is transformed to a problem of finding the smallest real eigenvalue of a quadratic eigenvalue problem, called QEPmin (Section 2.3). Our major contributions are as follows:

(i) Although transforming CRQopt into LGopt and QEPmin is not really new, our formulations of LGopt and QEPmin set them up onto a natural path for use in Krylov subspace type projection methods that only requires matrix-vector products. Therefore, the formulations are suitable for large scale CRQopt. We rigorously prove the equivalences among the three problems while they are only loosely argued previously as, see for example [10]. As far as subtle technicalities are concerned, we prove that the leftmost eigenvalue in the complex plane is real, which has a significant implication when it comes to numerical computations.

(ii) We devise a Lanczos algorithm to solve the induced optimization problems: LGopt and QEPmin. This algorithm is made possible, as we argued moments ago, by our different formulations from what in literature. Along the way, we also propose an efficient numerical algorithm for the type of secular equations arising from solving each projected LGopt. We establish a quantitative convergence analysis for the Lanczos algorithm and obtain error bounds on approximations generated by the algorithm. These error bounds are in general sharp in the worst case as demonstrated by artificially designed numerical examples.

(iii) We apply the proposed Lanczos algorithm to large scale CRQopt for constrained clustering, an extension of the well-known spectral algorithm with linear constraints to encode prior knowledge labels. We observe that the new Lanczos algorithm is 2 to 23 times faster than FAST-GE-2.0 [18] for constrained image segmentation.

1.3 Organization and notation

The rest of the article is organized as follows. In Section 2, we investigate the theoretical aspects of CRQopt (1.1) such as the feasible set, existence of a minimizer, and transforming CRQopt (1.1) into equivalent optimization problems with rigorous justifications. A Krylov subspace projection approach for solving CRQopt (1.1) via its equivalent optimization problems are detailed in Section 3. The convergence analysis is given in Section 3.5. Numerical examples to demonstrate the sharpness of convergence estimates are presented in Section 3.6. Section 4 describes an application of our algorithms to the constrained image segmentation problem. Concluding remarks are in Section 5. There are three appendices. Appendix A explains how to solve the secular equation arising from solving the LGopt. Appendix B proves the equivalence between CRQopt (1.1) and an eigenvalue optimization problem proposed by Eriksson, Olsson and Kahl [6]. Appendix C documents CRQPACK, a software package for an implementation of Lanczos algorithm and reproduce numerical experiments presented in this paper.

Throughout the article, $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ are sets of real numbers, columns vectors of dimension n, and $m\times n$ matrices, respectively. $\mathbb{C}$, $\mathbb{C}^n$ and $\mathbb{C}^{m\times n}$ are sets of complex numbers, columns vectors of dimension n, and $m\times n$ matrices, respectively. We use MATLAB-like notation $X_{(i:j,k:l)}$ to denote the submatrix of X consisting of the intersections of rows i to j and columns k to l, and when $i:j$ is replaced by :, it means all rows, similarly for columns. For a vector $v\in\mathbb{C}$, $v_{(k)}$ refers the kth entry of v and $v_{(i:j)}$ is the subvector of v consisting of the ith to jth entries inclusive. An $n\times n$ identity matrix is I_n or simply I if its size is clear from the context, and e_j is the jth column of an identity matrix whose size is determined by the context. $\operatorname{diag}(c_1,c_2,\cdots,c_n)$ is an $n\times n$ diagonal matrix with diagonal elements $c_1,c_2,\cdots,c_n$. The imaginary unit is $\mathtt{i}=\sqrt{-1}$. For $X\in\mathbb{C}^{m\times n}$, X^{T}, $\mathcal{R}(X)$ and $\mathcal{N}(X)$ denote its transpose, range and null space, respectively. For a real symmetric matrix H, $\operatorname{eig}(H)$ stands for the set of all eigenvalues of H, and $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ denote the smallest and largest eigenvalue of H, respectively. $\|\cdot\|_p$ $(1\leqslant p\leqslant\infty)$ is the ℓ_p-vector or ℓ_p-operator norm, respectively, depending on the argument. As a special case, $\|\cdot\|_2$ or $\|\cdot\|$ is either the Euclidean norm of vector or the spectral norm of a matrix.

2 Theory

2.1 Feasible set and solution existence

Let n_0 be the unique minimal norm solution of $C^{\mathrm{T}}v=b$:

$$n_0=(C^{\mathrm{T}})^{\dagger}b, \tag{2.1}$$

where $X^{\dagger}$ is the Moore-Penrose inverse of X [1,4,38]. By the assumption of $\operatorname{rank}(C)=m$, $C^{\dagger}=(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}$ and $(C^{\mathrm{T}})^{\dagger}=(C^{\dagger})^{\mathrm{T}}=C(C^{\mathrm{T}}C)^{-1}$. The most important orthogonal projection throughout this article is

$$P=I-CC^{\dagger}, \tag{2.2}$$

which orthogonally projects any vector onto $\mathcal{N}(C^{\mathrm{T}})$, the null space of C^{T} [38]. Any $v\in\mathbb{R}^n$ that satisfies $C^{\mathrm{T}}v=b$ can be orthogonally decomposed as

$$v=(I-P)v+Pv=n_0+Pv\in n_0+\mathcal{N}(C^{\mathrm{T}}). \tag{2.3}$$

Evidently $\|v\|^2=\|n_0\|^2+\|Pv\|^2$, which, together with the unit length constraint (1.1b), lead to the following immediate conclusions about the solvability of CRQopt (1.1):

- If $\|n_0\| > 1$, then there is no unit vector v satisfying $C^{\mathrm{T}}v = b$. This is due to the fact that any v satisfying $C^{\mathrm{T}}v = b$ has norm no smaller than $\|n_0\|$. Thus CRQopt (1.1) has no solution.

- If $\|n_0\| = 1$, then $v = n_0$ is the only unit vector that satisfies $C^{\mathrm{T}}v = b$. Thus CRQopt (1.1) has a unique minimizer $v = n_0$.

- If $\|n_0\| < 1$, then there are infinitely many feasible vectors v that satisfy $C^{\mathrm{T}}v = b$.

Therefore only the case $\|n_0\| < 1$ needs further investigation. Consequently, throughout the rest of the article, we will assume $\|n_0\| < 1$.

2.2 Equivalent LGopt

Using the orthogonal decomposition (2.3), we have

$$v^{\mathrm{T}}Av = v^{\mathrm{T}}PAPv + 2v^{\mathrm{T}}PAn_0 + n_0^{\mathrm{T}}An_0, \tag{2.4a}$$

$$v^{\mathrm{T}}v = \|n_0\|^2 + \|Pv\|^2. \tag{2.4b}$$

Since $n_0^{\mathrm{T}}An_0$ and $\|n_0\|$ are constants, CRQopt (1.1) is equivalent to the following constrained quadratic minimization problem:

$$\text{CQopt:} \quad \begin{cases} \min \; v^{\mathrm{T}}PAPv + 2v^{\mathrm{T}}b_0, & (2.5a) \\ \text{ s.t. } \|Pv\| = \gamma, & (2.5b) \\ \qquad v \in n_0 + \mathcal{N}(C^{\mathrm{T}}), & (2.5c) \end{cases}$$

where

$$b_0 = PAn_0 \in \mathcal{N}(C^{\mathrm{T}}), \quad \gamma := \sqrt{1 - \|n_0\|^2} > 0. \tag{2.6}$$

Necessarily, $0 < \gamma < 1$. However, in the rest of our development, unless we refer back to CRQopt (1.1), $\gamma < 1$ can be removed, i.e., γ can be any positive number.

Theorem 2.1. *v_* is a minimizer of* CRQopt (1.1) *if and only if v_* is a minimizer of* CQopt (2.5).

One way to solve CQopt (2.5) is the method of the Lagrangian multipliers. It seeks the stationary points of the Lagrangian function

$$\mathscr{L}(v,\lambda) = v^{\mathrm{T}}PAPv + 2v^{\mathrm{T}}b_0 - \lambda(v^{\mathrm{T}}Pv - \gamma^2). \tag{2.7}$$

Differentiating $\mathscr{L}$ with respect to v and λ, we get

$$(PA - \lambda I)Pv = -b_0, \tag{2.8a}$$

$$\|Pv\| = \gamma. \tag{2.8b}$$

Let $u=Pv\in\mathcal{N}(C^{\mathrm{T}})$. Then $u=Pu$ and $v=n_0+u$. The Lagrangian equations in (2.8) are equivalent to the following equations:

$$(PAP-\lambda I)u=-b_0, \tag{2.9a}$$
$$\|u\|=\gamma, \tag{2.9b}$$
$$u\in\mathcal{N}(C^{\mathrm{T}}). \tag{2.9c}$$

In fact, any solution (λ,v) of (2.8) gives rise to a solution (λ,u) with $u=Pv$ of (2.9), and conversely any solution (λ,u) of (2.9) leads to a solution (λ,v) with $v=n_0+u$ of (2.8).

The system of equations (2.9) has more than one solution pairs (λ,u) since CQopt (2.5) is non-convex and any local minimum or maximum has a corresponding solution pair (λ,u) of (2.9). In addition, in Section 2.3, we will show that under some conditions, an eigenpair of a quadratic eigenvalue problem (QEP) leads to a solution of (2.9), and the number of eigenpairs is not unique. We seek a pair (λ,u) of (2.9) that minimizes the objective function of (2.5) for $v\in\mathbb{R}^n$. Note that

$$\begin{aligned} f(v) &:= v^{\mathrm{T}}PAPv+2v^{\mathrm{T}}b_0=v^{\mathrm{T}}PAPv+2v^{\mathrm{T}}PAn_0 \\ &\overset{u=Pv}{=} u^{\mathrm{T}}Au+2u^{\mathrm{T}}An_0 \overset{u=Pu}{=} u^{\mathrm{T}}PAPu+2u^{\mathrm{T}}PAn_0 \\ &= u^{\mathrm{T}}PAPu+2u^{\mathrm{T}}b_0=f(u), \end{aligned} \tag{2.10}$$

i.e., $f(v)=f(u)$ for $v\in\mathbb{R}^n$ and $u=Pv$. Therefore minimizing $f(v)$ over $v\in\mathbb{R}^n$ is equivalent to minimizing $f(u)$ over $u\in\mathcal{N}(C^{\mathrm{T}})$. The following lemma compares the value of f at different solution pairs (λ,u) of (2.9). The proof of the lemma is inspired by Gander [9] on solving a least squares problem with a quadratic constraint.

Lemma 2.1. *For two solution pairs (λ_i,u_i) for $i=1,2$ of the Lagrangian system of equations* (2.9), *$\lambda_1<\lambda_2$ if and only if $f(u_1)<f(u_2)$.*

Proof. The proof relies on the following three facts: (i) For any solution pair (λ,u) of (2.9), we have

$$\lambda u=PAPu+b_0 \quad\Rightarrow\quad \lambda=\frac{1}{u^{\mathrm{T}}u}u^{\mathrm{T}}(PAPu+b_0)=\frac{1}{\gamma^2}u^{\mathrm{T}}(PAPu+b_0). \tag{2.11}$$

(ii) Given (λ_i,u_i) for $i=1,2$, satisfying (2.9), we have

$$\begin{aligned} f(u_1) &= u_1^{\mathrm{T}}PAPu_1+2u_1^{\mathrm{T}}b_0 \overset{(2.9a)}{=} -b_0^{\mathrm{T}}u_1+\lambda_1u_1^{\mathrm{T}}u_1+2u_1^{\mathrm{T}}b_0 \\ &\overset{(2.9b)}{=} u_1^{\mathrm{T}}b_0+\lambda_1\gamma^2 \overset{(2.9a)}{=} -u_2^{\mathrm{T}}(PAP-\lambda_2I)u_1+\lambda_1\gamma^2. \end{aligned}$$

Similarly, we have $f(u_2)=-u_1^{\mathrm{T}}(PAP-\lambda_1I)u_2+\lambda_2\gamma^2$. Therefore

$$f(u_1)-f(u_2)=(\lambda_1-\lambda_2)(\gamma^2-u_1^{\mathrm{T}}u_2). \tag{2.12}$$

(iii) For u_i of norm γ, by the Cauchy-Schwartz inequality, $u_1^{\mathrm{T}}u_2 \leqslant \|u_1\|\,\|u_2\| = \gamma^2$, and $u_1^{\mathrm{T}}u_2 = \|u_1\|\,\|u_2\| = \gamma^2$ if and only if $u_1 = u_2$. Hence if $u_1 \neq u_2$, then $\gamma^2 - u_1^{\mathrm{T}}u_2 > 0$.

Now we are ready to prove the claim of the lemma. If $\lambda_1 < \lambda_2$, then $u_1 \neq u_2$ otherwise (2.11) would imply $\lambda_1 = \lambda_2$, and thus $f(u_1) < f(u_2)$ by (2.12). On the other hand, if $f(u_1) < f(u_2)$, then $\gamma^2 - u_1^{\mathrm{T}}u_2 > 0$ because $\gamma^2 - u_1^{\mathrm{T}}u_2 \geqslant 0$ always and it cannot be 0 by (2.12), and thus $\lambda_1 - \lambda_2 < 0$ again by (2.12). □

As a consequence of Lemma 2.1, we find that solving CQopt (2.5) is equivalent to solving the smallest Lagrangian multiplier λ of (2.7), i.e., those λ that satisfy (2.9). Specifically, solving CQopt (2.5) is equivalent to solving the following Lagrangian minimization problem:

$$\text{LGopt:} \quad \begin{cases} \min\ \lambda & \text{(2.13a)} \\ \quad \text{s.t. } (PAP - \lambda I)u = -b_0, & \text{(2.13b)} \\ \qquad \|u\| = \gamma, & \text{(2.13c)} \\ \qquad u \in \mathcal{N}(C^{\mathrm{T}}). & \text{(2.13d)} \end{cases}$$

Theorem 2.2. *If v_* is a minimizer of* CQopt (2.5)*, then (λ_*, u_*) with $u_* = Pv_*$ and $\lambda_* = \frac{1}{\gamma^2}u_*^{\mathrm{T}}(PAPu_* + b_0)$ is a minimizer of* LGopt (2.13)*. Conversely if (λ_*, u_*) is a minimizer of* LGopt (2.13)*, then $v_* = n_0 + u_*$ is a minimizer of* CQopt (2.5)*.*

The case $b_0 = PAn_0 = 0$, which includes but is not equivalent to the homogeneous CRQopt (1.1) (i.e., $b = 0$) treated in [14,15] can be dealt with as follows. Suppose $b_0 = 0$ and let θ_1 be the smallest eigenvalue of PAP. Keep in mind that PAP always has an eigenvalue 0 with multiplicity m associated with the subspace $\mathcal{N}(C^{\mathrm{T}})^{\perp} = \mathcal{R}(C)$, the column space of C. There are the following two subcases:

- **Subcase $\theta_1 \neq 0$:** Then[†] $\theta_1 < 0$. Let z_1 be a corresponding eigenvector of PAP. Then $z_1 = PAPz_1/\theta_1 \in \mathcal{N}(C^{\mathrm{T}})$. So (θ_1, z_1) is a minimizer of LGopt (2.13) and therefore z_1 is a minimizer of CQopt (2.5), which in turn implies that $v_* = n_0 + \gamma z_1/\|z_1\|$ is a minimizer of CRQopt (1.1).

- **Subcase $\theta_1 = 0$:** If there exists a corresponding eigenvector $z_1 \in \mathcal{N}(C^{\mathrm{T}})$, i.e., $Pz_1 \neq 0$, then (θ_1, Pz_1) is a minimizer of LGopt (2.13) and therefore Pz_1 is a minimizer of CQopt (2.5), which in turn implies that $v_* = n_0 + \gamma Pz_1/\|Pz_1\|$ is a minimizer of CRQopt (1.1). Otherwise there exists no corresponding eigenvector z_1 such that $Pz_1 \neq 0$. Let θ_2 be the second smallest eigenvalue of PAP, which is nonzero, and z_2 a corresponding eigenvector. Then $z_2 = PAPz_2/\theta_2 \in \mathcal{N}(C^{\mathrm{T}})$, and (θ_2, z_2) is a minimizer of LGopt (2.13) and therefore z_2 is a minimizer of CQopt (2.5), which in turn implies that $v_* = n_0 + \gamma z_2/\|z_2\|$ is a minimizer of CRQopt (1.1).

In view of such a quick resolution for the case $b_0 = 0$, in the rest of this article, we will assume

$$b_0 = PAn_0 \neq 0. \tag{2.14}$$

[†] This cannot happen if A is positive semidefinite.

2.3 Equivalent QEPmin

Let (λ,u) be a feasible pair of LGopt (2.13) and $\lambda\notin\mathrm{eig}(PAP)$. From (2.13b), we can write $u=-(PAP-\lambda I)^{-1}b_0$, and then

$$\gamma^2=u^{\mathrm{T}}u=b_0^{\mathrm{T}}(PAP-\lambda I)^{-2}b_0=b_0^{\mathrm{T}}z, \tag{2.15}$$

where $z=(PAP-\lambda I)^{-2}b_0$, or equivalently, $(PAP-\lambda I)^2z=b_0$. Therefore $b_0^{\mathrm{T}}z/\gamma^2=1$ by (2.15), and the pair (λ,z) satisfies the quadratic eigenvalue problem (QEP):

$$(PAP-\lambda I)^2z=b_0=b_0\cdot 1=b_0\left(b_0^{\mathrm{T}}z/\gamma^2\right)=\frac{1}{\gamma^2}b_0b_0^{\mathrm{T}}z. \tag{2.16}$$

We claim that any z satisfying (2.16) is in $\mathcal{N}(C^{\mathrm{T}})$. To see this, we expand $(PAP-\lambda I)^2z$ and extract λ^2z from $(PAP-\lambda I)^2z=b_0$ to get

$$z=\frac{1}{\lambda^2}\left[-(PAP)^2z+2\lambda\cdot PAPz+b_0\right]\in\mathcal{N}(C^{\mathrm{T}}),$$

where we have used the assumption $\lambda\notin\mathrm{eig}(PAP)$ to conclude $\lambda\neq 0$, and $b_0=PAn_0\in\mathcal{N}(C^{\mathrm{T}})$. Therefore we have shown that under the assumption that LGopt (2.13) has no feasible pair (λ,u) with $\lambda\in\mathrm{eig}(PAP)$, any feasible pair (λ,u) of LGopt (2.13) satisfies QEP (2.16) with $z\in\mathcal{N}(C^{\mathrm{T}})$.

Next, we prove that any pair (λ,z) satisfying

$$0\neq z\in\mathcal{N}(C^{\mathrm{T}}),\quad \lambda\notin\mathrm{eig}(PAP)\ \text{ and }\ \text{QEP (2.16)}, \tag{2.17}$$

leads to a feasible pair of the Lagrange equations (2.13). First we note that $b_0^{\mathrm{T}}z\neq 0$; otherwise we would have $(PAP-\lambda I)^2z=0$ by (2.16), implying $z=0$ since $\lambda\notin\mathrm{eig}(PAP)$, a contradiction. Let (λ,z) be a scalar-vector pair satisfying (2.17). Define $u:=-(PAP-\lambda I)^{-1}b_0$. Then $(PAP-\lambda I)u=-b_0$, i.e., (2.13b) holds, and also

$$\lambda u=PAPu+b_0\quad\Rightarrow\quad u=\frac{1}{\lambda}(PAPu+b_0)\in\mathcal{N}(C^{\mathrm{T}}),$$

i.e., (2.13d) holds. Without loss of generality, we may scale z such that $b_0^{\mathrm{T}}z=\gamma^2$. It follows from (2.16) that

$$(PAP-\lambda I)^2z=b_0\quad\Rightarrow\quad z=(PAP-\lambda I)^{-2}b_0,$$

implying

$$1=\frac{1}{\gamma^2}b_0^{\mathrm{T}}z=\frac{1}{\gamma^2}b_0^{\mathrm{T}}(PAP-\lambda I)^{-2}b_0=\frac{1}{\gamma^2}u^{\mathrm{T}}u\quad\Rightarrow\quad\|u\|=\gamma,$$

i.e., (2.13c) holds. Lemma 2.2 summarizes what we have just proved.

Lemma 2.2. *Suppose the constraints of* LGopt (2.13) *has no feasible pair* (λ,u) *with* $\lambda\in\operatorname{eig}(PAP)$*, and suppose that* QEP (2.16) *has no solution pair* (λ,z) *with* $0\neq z\in\mathcal{N}(C^{\mathrm{T}})$ *and* $\lambda\in\operatorname{eig}(PAP)$*. Then any pair* (λ,u) *satisfying the constraints of* LGopt (2.13) *gives rise to a pair* (λ,z) *with* $z=(PAP-\lambda I)^{-2}b_0$ *that satisfies* QEP (2.16)*. Conversely, any pair* (λ,z) *with* $z\neq 0$ *satisfying* QEP (2.16) *leads to a pair* (λ,u) *with* $u:=-(PAP-\lambda I)^{-1}b_0$ *that satisfies the constraints of* LGopt (2.13)*.*

As a corollary of Lemma 2.2, we conclude that LGopt (2.13) is equivalent to

$$\text{QEPmin:}\quad\begin{cases}\min\ \lambda & \text{(2.18a)}\\ \quad\text{s.t. } (PAP-\lambda I)^2z=\gamma^{-2}b_0b_0^{\mathrm{T}}z, & \text{(2.18b)}\\ \qquad\quad\lambda\in\mathbb{R},\ 0\neq z\in\mathcal{N}(C^{\mathrm{T}}), & \text{(2.18c)}\end{cases}$$

under the assumptions of Lemma 2.2. Soon we show that LGopt (2.13) and QEPmin (2.18) are still equivalent even without the assumptions.

We name the minimization problem (2.18) QEPmin because the constraint (2.18b) is a quadratic eigenvalue problem (QEP). Although this QEP generally may have complex eigenvalues λ, the "min" in (2.18a) implicitly restricts the consideration only to the real eigenvalues λ of QEP (2.18b) in the context of QEPmin (2.18). In this sense, there is no need to specify $\lambda\in\mathbb{R}$ in (2.18c), but we are doing it anyway to emphasize the implication. This comment applies to two other minimization problems pQEPmin (2.26) and rQEPmin (3.22) later that involve a QEP as a constraint as well.

In the rest of this section, we prove the equivalence between LGopt (2.13) and QEPmin (2.18) without the assumptions of Lemma 2.2. The key idea is to remove the null space conditions $u,z\in\mathcal{N}(C^{\mathrm{T}})$ by projecting Eqs. (2.13b), (2.13c) in LGopt and (2.18b) in QEPmin onto an appropriate subspace.

2.4 pLGopt

Let $S=[S_1,\ S_2]\in\mathbb{R}^{n\times n}$ be an orthogonal matrix with

$$\mathcal{R}(S_1)=\mathcal{N}(C^{\mathrm{T}}),\quad \mathcal{R}(S_2)=\mathcal{N}(C^{\mathrm{T}})^{\perp}. \tag{2.19}$$

Since $\operatorname{rank}(C)=m$, we know $S_1\in\mathbb{R}^{n\times(n-m)}$ and $S_2\in\mathbb{R}^{n\times m}$. It can be verified that the projection matrix $P=I-CC^{\dagger}$ in (2.2) can be written as

$$P=S_1S_1^{\mathrm{T}}=I-S_2S_2^{\mathrm{T}}, \tag{2.20}$$

and

$$PS_1=S_1,\quad PS_2=0.$$

Set

$$g_0=S_1^{\mathrm{T}}b_0,\quad H=S_1^{\mathrm{T}}PAPS_1=S_1^{\mathrm{T}}AS_1\in\mathbb{R}^{(n-m)\times(n-m)}, \tag{2.21}$$

we have

$$S^{\mathrm{T}}PAPS=\begin{bmatrix} S_1^{\mathrm{T}}PAPS_1 & S_1^{\mathrm{T}}PAPS_2 \\ S_2^{\mathrm{T}}PAPS_1 & S_2^{\mathrm{T}}PAPS_2 \end{bmatrix}=\begin{array}{c|cc} & {\scriptstyle n-m} & {\scriptstyle m} \\ {\scriptstyle n-m} & H & 0 \\ {\scriptstyle m} & 0 & 0 \end{array}, \tag{2.22a}$$

$$S^{\mathrm{T}}b_0=\begin{bmatrix} S_1^{\mathrm{T}}b_0 \\ S_2^{\mathrm{T}}b_0 \end{bmatrix}=\begin{array}{c|c} {\scriptstyle n-m} & g_0 \\ {\scriptstyle m} & 0 \end{array}. \tag{2.22b}$$

Immediately from the decomposition (2.22a), we conclude the following lemma.

Lemma 2.3. *The eigenvalues of PAP consist of those of H and 0 with multiplicities m, i.e., $\operatorname{eig}(PAP)=\operatorname{eig}(H)\cup\{0,0,\cdots,0\}$. If $0\neq\lambda\in\operatorname{eig}(PAP)$, then $\lambda\in\operatorname{eig}(H)$ and its associated eigenvector must be in $\mathcal{N}(C^{\mathrm{T}})$. The matrix PAP has more than m eigenvalues 0 if and only if H is singular. For each eigenvalue 0 of PAP coming from $\operatorname{eig}(H)$, there is an eigenvector z of PAP such that $Pz\neq 0$ (in fact, Pz is an eigenvector for that particular eigenvalue 0 as well).*

To explicitly eliminate the constraint $u\in\mathcal{N}(C^{\mathrm{T}})$ in LGopt (2.13), we project LGopt (2.13) onto $\mathcal{R}(S_1)$ and introduce the following projected minimization problem

$$\text{pLGopt:}\quad\begin{cases} \min\ \lambda & (2.23a)\\ \text{s.t. } (H-\lambda I)y=-g_0, & (2.23b)\\ \qquad \|y\|=\gamma. & (2.23c)\end{cases}$$

The next theorem establishes the equivalence between LGopt (2.13) and pLGopt (2.23).

Theorem 2.3. *The pair (λ_*,y_*) is a minimizer of* pLGopt (2.23) *if and only if (λ_*,u_*) with $u_*=S_1y_*$ is a minimizer of* LGopt (2.13).

Proof. We begin by showing the equivalence between the constraints of LGopt (2.13) and those of pLGopt (2.23). Note that any $0\neq u\in\mathcal{N}(C^{\mathrm{T}})$ can be expressed by $u=S_1y$ for some $0\neq y\in\mathbb{R}^{n-m}$ and vice versa. Making use of (2.22), we have

$$\begin{aligned} S^{\mathrm{T}}[(PAP-\lambda I)u+b_0]&=S^{\mathrm{T}}(PAP-\lambda I)SS^{\mathrm{T}}u+S^{\mathrm{T}}b_0\\ &=\begin{bmatrix} H-\lambda I & 0\\ 0 & -\lambda I\end{bmatrix}\begin{bmatrix} y\\ 0\end{bmatrix}+\begin{bmatrix} g_0\\ 0\end{bmatrix} \end{aligned}\tag{2.24}$$

and

$$u^{\mathrm{T}}u=y^{\mathrm{T}}S_1^{\mathrm{T}}S_1y=y^{\mathrm{T}}y. \tag{2.25}$$

Now if (λ,u) satisfies the constraints of LGopt (2.13), then $S^{\mathrm{T}}[(PAP-\lambda I)u+b_0]=0$ because of (2.13b), $u=S_1y$ for some y because of (2.13d), and $\|y\|=\gamma$ because of (2.13c) and (2.25). It follows from (2.24) that $(H-\lambda I)y+g_0=0$. Thus (λ,y) satisfies the constraints of pLGopt (2.23).

On the other hand, suppose (λ,y) satisfies the constraints of pLGopt (2.23). Let $u=S_1y\in\mathcal{N}(C^{\mathrm{T}})$. Both (2.24) and (2.25) remain valid. Then $S^{\mathrm{T}}[(PAP-\lambda I)u+b_0]=0$ which implies $(PAP-\lambda I)u+b_0=0$ because S^{T} is an orthogonal matrix. Also $\|u\|=\gamma$ by (2.25). This completes the proof of that (λ,u) satisfies the constraints of LGopt (2.13).

Therefore, LGopt (2.13) and pLGopt (2.23) have the same optimal value λ_*. More than that, if (λ_*,u_*) is a minimizer of LGopt (2.13), then there exists y_* such that $u_*=S_1y_*$ and that (λ_*,y_*) is a minimizer of pLGopt (2.23), and vice versa. □

We note that for a modest-sized CRQopt (1.1), say n up to 2000, we may as well perform the reduction to form pLGopt (2.23) explicitly. Due to its modest size, pLGopt (2.23) can be solved as a dense matrix computational problem. The detail is buried later in the proof of Lemma 2.4.

2.5 pQEPmin

For the same purpose as we projected the Lagrange equations, we introduce the following projected minimization problem as the counterpart of QEPmin (2.18):

$$\text{pQEPmin:}\quad\begin{cases}\min\lambda & \text{(2.26a)}\\ \quad\text{s.t. } (H-\lambda I)^2w=\gamma^{-2}g_0g_0^{\mathrm{T}}w, & \text{(2.26b)}\\ \qquad\lambda\in\mathbb{R},\ w\neq 0. & \text{(2.26c)}\end{cases}$$

The equation in (2.26b) has an appearance of a QEP. As stated, the optimal value of pQEPmin (2.26) is the smallest real eigenvalue of QEP (2.26b). The next theorem establishes the equivalence between QEPmin (2.18) and pQEPmin (2.26).

Theorem 2.4. *The pair* (λ_*,w_*) *is a minimizer of* pQEPmin (2.26) *if and only if* (λ_*,z_*) *with* $z_*=S_1w_*$ *is a minimizer of* QEPmin (2.18).

Proof. We begin by showing the equivalence between the constraints of QEPmin (2.18) and those of pQEPmin (2.26). Keeping (2.22) in mind, we have for any $z=S_1w$

$$\begin{aligned}&S^{\mathrm{T}}\Big[(PAP-\lambda I)^2z-\gamma^{-2}b_0b_0^{\mathrm{T}}z\Big]\\&=S^{\mathrm{T}}(PAP-\lambda I)SS^{\mathrm{T}}(PAP-\lambda I)SS^{\mathrm{T}}z-\gamma^{-2}S^{\mathrm{T}}b_0b_0^{\mathrm{T}}SS^{\mathrm{T}}z\\&=\begin{bmatrix}(H-\lambda I)^2 & 0\\ 0 & \lambda^2I\end{bmatrix}\begin{bmatrix}w\\0\end{bmatrix}-\begin{bmatrix}\gamma^{-2}g_0g_0^{\mathrm{T}} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}w\\0\end{bmatrix}.\end{aligned}\tag{2.27}$$

Now if (λ,z) satisfies the constraints of QEPmin (2.18), then $0\neq z\in\mathcal{N}(C^{\mathrm{T}})$ and thus $z=S_1w$ for some $0\neq w\in\mathbb{R}^{n-m}$. Therefore, by (2.27), (λ,w) satisfies (2.26b).

On the other hand, suppose (λ,w) satisfies (2.26b) and (2.26c). Let $z=S_1w\in\mathcal{N}(C^{\mathrm{T}})$. Then $z\neq 0$ and by (2.27), $S^{\mathrm{T}}[(PAP-\lambda I)^2z-\gamma^{-2}b_0b_0^{\mathrm{T}}z]=0$. Since S^{T} is orthogonal, we get (2.18b). This proves that (λ,z) satisfies the constraints of QEPmin (2.18).

Therefore, QEPmin (2.18) and pQEPmin (2.26) have the same optimal value λ_*. More than that, if (λ_*,z_*) is a minimizer of QEPmin (2.18), then there exists $w_*\neq 0$ such that $z_*=S_1w_*$ and that (λ_*,w_*) is a minimizer of pQEPmin (2.26), and vice versa. □

2.6 The equivalence of pLGopt and pQEPmin

Although, in leading to pLGopt (2.23) and pQEPmin (2.26), the matrix H and the vector g_0 are derived from reducing A, C, and b in the original CRQopt (1.1), the developments in this section does not require that. Given this, in the rest of this section, we consider general pLGopt (2.23) and pQEPmin (2.26) with[‡]

$$H\in\mathbb{R}^{\ell\times\ell},\quad H^{\mathrm{T}}=H,\quad 0\neq g_0\in\mathbb{R}^{\ell},\ \text{ and }\ \gamma>0.$$

To set up the stage for the rest of this subsection, let $H=Y\Theta Y^{\mathrm{T}}$ be the eigen-decomposition of H:

$$H=Y\Theta Y^{\mathrm{T}}\ \text{ with }\ \Theta=\operatorname{diag}(\theta_1,\theta_2,\cdots,\theta_\ell),\quad Y=[y_1,y_2,\cdots,y_\ell],\quad Y^{\mathrm{T}}Y=I_\ell. \tag{2.28}$$

Without loss of generality, we arrange θ_i in the ascending order:

$$\theta_1=\theta_2=\cdots=\theta_d<\theta_{d+1}\leqslant\cdots\leqslant\theta_\ell,$$

and set $\lambda_{\min}(H)=\theta_1$. Define the secular function

$$\chi(\lambda):=g_0^{\mathrm{T}}(H-\lambda I)^{-2}g_0-\gamma^2=(Y^{\mathrm{T}}g_0)^{\mathrm{T}}(\Theta-\lambda I)^{-2}(Y^{\mathrm{T}}g_0)-\gamma^2=\sum_{i=1}^{l}\frac{\xi_i^2}{(\lambda-\theta_i)^2}-\gamma^2, \tag{2.29}$$

where $\xi_i=g_0^{\mathrm{T}}y_i$ for $i=1,2,\cdots,n$, and let

$$j_0=\min\{i:\ \xi_i\neq 0\}. \tag{2.30}$$

Lemma 2.4. *Let (λ_*,y_*) be a minimizer of* pLGopt (2.23). *The following statements hold.*

(a) $\lambda_*\leqslant\lambda_{\min}(H)$.

(b) $\lambda_*=\lambda_{\min}(H)$ *if and only if*

$$g_0\perp\mathcal{U}\ \textit{ and }\ \|(H-\lambda_{\min}(H)I)^{\dagger}g_0\|_2\leqslant\gamma,$$

where $\mathcal{U}$ is the eigenspace of H associated with its eigenvalue $\lambda_{\min}(H)$.

(c) *If $g_0\not\perp\mathcal{U}$, then $\lambda_*<\lambda_{\min}(H)$ and λ_* is the smallest root of the secular function $\chi(\lambda)$, and $y_*=-(H-\lambda_*I)^{-1}g_0$.*

[‡] Unlike before, there is no need to assume $\gamma<1$. In addition, the size of square matrix H and vector g_0 can be arbitrary, not necessarily equal to $n-m$.

Proof. The secular function $\chi(\lambda)$ in (2.29) is continuous on $(-\infty,\theta_1)$ and $\lim_{\lambda\to-\infty}\chi(\lambda)=-\gamma^2<0$. Since

$$\chi'(\lambda)=-2\sum_{i=1}^{\ell}\frac{\xi_i^2}{(\lambda-\theta_i)^3}>0 \quad \text{for } \lambda<\theta_1\,,$$

$\chi(\lambda)$ is strictly increasing in $(-\infty,\theta_1)$. We have the following situations to deal with:

(1) If $g_0\not\perp\mathcal{U}$, then $\sum_{i=1}^{d}\xi_i^2>0$, i.e., $j_0\leqslant d$, then $\lim_{\lambda\to\theta_1^-}\chi(\lambda)=+\infty>0$. There exists a unique $\lambda_*\in(-\infty,\theta_1)$ such that $\chi(\lambda_*)=0$. Let $y_*=-(H-\lambda_*I)^{-1}g_0$, then

$$y_*^{\mathrm{T}}y_*=g_0^{\mathrm{T}}(H-\lambda_*I)^{-2}g_0=\chi(\lambda_*)+\gamma^2=\gamma^2.$$

Therefore, (λ_*,y_*) satisfies the constraints of pLGopt (2.23).

(2) Suppose that $g_0\perp\mathcal{U}$, then $\sum_{i=1}^{d}\xi_i^2=0$, i.e., $j_0>d$. Let

$$w=-(H-\theta_1I)^{\dagger}g_0=-\sum_{i=d+1}^{\ell}\frac{\xi_i}{\theta_i-\theta_1}y_i.$$

Then $(H-\theta_1I)w=-g_0$ and $\lim_{\lambda\to\theta_1^-}\chi(\lambda)=w^{\mathrm{T}}w-\gamma^2$. There are the following three subcases:

(i) If $\|w\|>\gamma$, then there exists a unique $\lambda_*\in(-\infty,\theta_1)$ such that $\chi(\lambda_*)=0$. Moreover (λ_*,y_*) with $y_*=-(H-\lambda_*I)^{-1}g_0$ satisfies the constraints of pLGopt (2.23).

(ii) If $\|w\|=\gamma$, then (λ_*,y_*) with $\lambda_*=\theta_1$ and $y_*=w$ satisfies the constraints of pLGopt (2.23).

(iii) If $\|w\|<\gamma$, then (λ_*,y_*) with $\lambda_*=\theta_1$ and $y_*=w+\sqrt{\gamma^2-\|w\|^2}y_1$ satisfies the constraints of pLGopt (2.23).

Hence we proved that (λ_*,y_*) satisfies the constraints of pLGopt (2.23) for all situations.

Now we prove λ_* is the smallest solution which satisfies the constraints of pLGopt (2.23). Suppose there exists $\widehat{\lambda}<\lambda_*$ such that $(\widehat{\lambda},\widehat{y})$ satisfies the constraints of pLGopt (2.23), then $\widehat{\lambda}<\lambda_*\leqslant\theta_1$, so $\widehat{\lambda}\notin\operatorname{eig}(H)$. Therefore, in order to make $(\widehat{\lambda},\widehat{y})$ satisfies (2.23b), we have $\widehat{y}=-(H-\widehat{\lambda}I)^{-1}g_0$. Note that $\lim_{\lambda\to\lambda_*^-}\chi(\lambda)\leqslant 0$ for all cases and $\chi(\lambda)$ is strictly increasing in $(-\infty,\lambda_*)$, so $\chi(\widehat{\lambda})=\widehat{y}^{\mathrm{T}}\widehat{y}-\gamma^2<0$, which is contradictory to (2.23c) that $\|\widehat{y}\|=\gamma$. Therefore, λ_* is the smallest Lagrangian multiplier, and thus (λ_*,y_*) is a minimizer of pLGopt (2.23).

For all situations, the smallest Lagrangian multiplier λ_* of pLGopt (2.23) satisfies $\lambda_*\leqslant\lambda_{\min}(H)$, as expected. Also $\lambda_*=\theta_1$ can only happen in the subcase (ii) or (iii). □

Buried in the proof above is a viable numerical algorithm to solve pLGopt (2.23), provided λ_* in the case (1) and the subcase (i) of the case (2) can be efficiently solved. In both cases, it is the unique root of secular equation $\chi(\lambda)=0$ in $(-\infty,\theta_1)$ in which $\chi(\lambda)$ monotonically increasing. A default method is Newton's method which applies the tangent line approximation, since both $\chi(\lambda)$ and its derivative $\chi'(\lambda)$ are rather straightforward to evaluate. However, this secular equation $\chi(\lambda)=0$ has a special rational form. Previous ideas in solving secular equations of similar types [2, 10, 21, 43] can be adopted to devise a much fast method than Newton's method. Details are presented in Appendix A.

Lemma 2.5. *If (λ,y) satisfies the constraints of* pLGopt (2.23)*, then there exists a vector $w\in\mathbb{R}^\ell$ such that (λ,w) satisfies the constraints of* pQEPmin (2.26)*. Specifically,*

$$w=\begin{cases}(H-\lambda I)^{-1}y, & \text{if } \lambda\notin\operatorname{eig}(H),\\ \text{the corresponding eigenvector of } H, & \text{if } \lambda\in\operatorname{eig}(H).\end{cases}$$

In particular, the optimal value of pQEPmin (2.26) *is less than or equal to the optimal value of* pLGopt (2.23)*.*

Proof. There are the cases to consider. (1) Case $\lambda\in\operatorname{eig}(H)$: Let w be an eigenvector of H corresponding to eigenvalue λ, i.e., $Hw=\lambda w$. By (2.23b), $g_0=-(H-\lambda I)y$, and thus

$$\gamma^{-2}g_0g_0^{\mathrm T}w=-\gamma^{-2}g_0y^{\mathrm T}(H-\lambda I)w=0.$$

Evidently, $(H-\lambda I)^2w=0$. Hence (λ,w) satisfies (2.26b). (2) Case $\lambda\notin\operatorname{eig}(H)$: Let $w=(H-\lambda I)^{-1}y$. Using (2.23b), we have

$$\begin{aligned}&(H-\lambda I)^2w=(H-\lambda I)y=-g_0,\\ &\gamma^{-2}g_0g_0^{\mathrm T}w=\gamma^{-2}g_0g_0^{\mathrm T}(H-\lambda I)^{-1}y=-\gamma^{-2}g_0y^{\mathrm T}y=-g_0.\end{aligned}$$

Again (λ,w) satisfies (2.26b).

Hence we proved that (λ,w) satisfies the constraints of pQEPmin (2.26). As a corollary, the optimal value of pQEPmin (2.26) is less than or equal to the optimal value of pLGopt (2.23). □

The next lemma claims a stronger conclusion than the last statement in the previous lemma.

Lemma 2.6. *The optimal value of* pLGopt (2.23) *is equal to the optimal value of* pQEPmin (2.26)*.*

Proof. Let (λ_*,y_*) be a minimizer of pLGopt (2.23), and let $\widehat{\lambda}$ be the optimal value of pQEPmin (2.26). By Lemma 2.5, we have $\widehat{\lambda}\leqslant\lambda_*$. It suffices to show that $\widehat{\lambda}<\lambda_*$ cannot happen. Assume, to the contrary, that $\widehat{\lambda}<\lambda_*$. By Lemma 2.4, we have $\widehat{\lambda}<\lambda_{\min}(H)$. In particular, $\widehat{\lambda}\notin\operatorname{eig}(H)$. Let $(\widehat{\lambda},\widehat{w})$ be a minimizer of pQEPmin (2.26). By (2.26b), we have

$$\frac{1}{\gamma^2}(\widehat{w}^{\mathrm T}g_0)^2=\widehat{w}^{\mathrm T}\frac{1}{\gamma^2}g_0g_0^{\mathrm T}\widehat{w}=\widehat{w}^{\mathrm T}(H-\widehat{\lambda}I)^2\widehat{w}>0,$$

implying $g_0^{\mathrm{T}}\widehat{w}\neq 0$. Let $\widehat{y}=-(\gamma^2/g_0^{\mathrm{T}}\widehat{w})(H-\widehat{\lambda}I)\widehat{w}$, and observe that

$$(H-\widehat{\lambda}I)\widehat{y}=-\frac{\gamma^2}{g_0^{\mathrm{T}}\widehat{w}}\cdot(H-\widehat{\lambda}I)^2\widehat{w}=-\frac{\gamma^2}{g_0^{\mathrm{T}}\widehat{w}}\cdot\gamma^{-2}g_0g_0^{\mathrm{T}}\widehat{w}=-g_0, \tag{2.31a}$$

$$\widehat{y}^{\mathrm{T}}\widehat{y}=\left(\frac{\gamma^2}{g_0^{\mathrm{T}}\widehat{w}}\right)^2\widehat{w}^{\mathrm{T}}(H-\widehat{\lambda}I)^2\widehat{w}=\left(\frac{\gamma^2}{g_0^{\mathrm{T}}\widehat{w}}\right)^2\frac{\widehat{w}^{\mathrm{T}}g_0g_0^{\mathrm{T}}\widehat{w}}{\gamma^2}=\gamma^2, \tag{2.31b}$$

i.e., $(\widehat{\lambda},\widehat{y})$ satisfies the constraints of pLGopt (2.23). This implies $\lambda_*\leqslant\widehat{\lambda}$, contradicting the assumption $\widehat{\lambda}<\lambda_*$. Therefore, $\widehat{\lambda}=\lambda_*$, as expected. □

We are ready to establish the equivalence between pLGopt (2.23) and pQEPmin (2.26).

Theorem 2.5 (The equivalence of pLGopt (2.23) **and pQEPmin** (2.26)**).**

(1) *Let* (λ_*,y_*) *be a minimizer of* pLGopt (2.23)*, then either* $\lambda_*<\lambda_{\min}(H)$ *or* $\lambda_*=\lambda_{\min}(H)$*, and there exists* w_* *such that* (λ_*,w_*) *is a minimizer of* pQEPmin (2.26). *Specifically,*

$$w_*=\begin{cases}(H-\lambda_*I)^{-1}y_*, & \text{if } \lambda_*<\lambda_{\min}(H),\\ \text{the corresponding eigenvector of } H, & \text{if } \lambda_*=\lambda_{\min}(H).\end{cases}$$

(2) *Conversely, if* (λ_*,w_*) *is a minimizer of* pQEPmin (2.26)*, then there exists* y_* *such that* (λ_*,y_*) *is a minimizer of* pLGopt (2.23). *Specifically,*

$$y_*=\begin{cases}-(\gamma^2/g_0^{\mathrm{T}}w_*)(H-\lambda_*I)w_*, & \text{if } g_0^{\mathrm{T}}w_*\neq 0,\\ x_*+\sqrt{\gamma^2-\|x_*\|^2}(w_*/\|w_*\|), & \text{if } g_0^{\mathrm{T}}w_*=0,\end{cases}$$

where $x_*=-(H-\lambda_*I)^{\dagger}g_0$ *in the case* $g_0^{\mathrm{T}}w_*=0$*, and it is guaranteed that* $\|x_*\|\leqslant\gamma$.

Proof. Item (1) is a consequence of Lemmas 2.5 and 2.6.

Consider item (2). Suppose (λ_*,w_*) is a minimizer of pQEPmin (2.26). By Lemma 2.6, it suffices to show that there exists y_* such that (λ_*,y_*) satisfies the constraints of pLGopt (2.23).

- Case $g_0^{\mathrm{T}}w_*\neq 0$: The equations in (2.31) hold with substitutions

  $$\widehat{\lambda}\to\lambda_*,\quad \widehat{y}\to y_*=-(\gamma^2/g_0^{\mathrm{T}}w_*)(H-\lambda_*I)w_*.$$

  So (λ_*,y_*) satisfies the constraints of pLGopt (2.23).

- Case $g_0^{\mathrm{T}}w_*=0$: By (2.26b), we find that $(H-\lambda_*I)^2w_*=0$, implying $(H-\lambda_*I)w_*=0$ since $H-\lambda_*I$ is real symmetric. Hence $\lambda_*\in\operatorname{eig}(H)$ and w_* is an associated eigenvector. Let x_* be the minimum norm solution of $(H-\lambda_*I)x_*=-g_0$.

  Note that we already know λ_* is the optimal value of pLGopt (2.23), which means there exists y such that (λ_*,y) satisfies (2.23b) and $\|y\|=\gamma$. On the other hand, x is minimal norm solution of (2.23b), so $\|x\|\leqslant\|y\|=\gamma$. Then it can be verified that (λ_*,y_*) with $y_*=x_*+\sqrt{\gamma^2-\|x_*\|^2}(w_*/\|w_*\|)$ satisfies the constraints of pLGopt (2.23).

This proves that (λ_*,y_*) satisfies the constraints of pLGopt (2.23). In addition, by Lemma 2.6, λ_* is the optimal value of pLGopt (2.23), which proves the result. □

The following theorem is about the uniqueness of the solution for pLGopt (2.23).

Theorem 2.6 (Uniqueness of the minimizer for pLGopt (2.23)). *Let (λ_*,w_*) be a minimizer of* pQEPmin (2.26).

(1) *If $g_0^{\mathrm{T}}w_* \neq 0$ for all possible minimizers for* pQEPmin (2.26), *then $\lambda_* < \lambda_{\min}(H)$ and the minimizer of* pLGopt (2.23) *is unique.*

(2) *If there exists a minimizer for* pQEPmin (2.26) *such that $g_0^{\mathrm{T}}w_* = 0$, then $\lambda_* = \lambda_{\min}(H)$ and the minimizer of* pLGopt (2.23) *is unique if and only if $\|x_*\| = \gamma$, where $x_* = -(H-\lambda_* I)^{\dagger}g_0$.*

Proof. (1) First we prove $\lambda_* < \lambda_{\min}(H)$. Suppose it is not true, i.e., $\lambda_* = \lambda_{\min}(H)$, let w_* be an eigenvector of H corresponding with eigenvalue $\lambda_{\min}(H)$, then by Theorem 2.5, (λ_*,w_*) is a minimizer of pQEPmin (2.26). Since QEP (2.26b) leads to $\gamma^{-2}g_0g_0^{\mathrm{T}}w_* = (H-\lambda_* I)^2 w_* = 0$ and $w_* \neq 0$, we have $g_0^{\mathrm{T}}w_* = 0$, which is contradictory to our assumption that $g_0^{\mathrm{T}}w_* \neq 0$ for all possible minimizers (λ_*,w_*) of pQEPmin (2.26). Therefore, $\lambda_* < \lambda_{\min}(H)$. In this case $(\lambda_*, x_* = -(H-\lambda_* I)^{-1}g_0)$ is the unique minimizer of pLGopt (2.23) since the $H-\lambda_* I$ is nonsingular and x_* is the unique solution of (2.23b).

(2) Making use of (2.26b), we have

$$(H-\lambda_* I)^2 w_* = \gamma^{-2}g_0g_0^{\mathrm{T}}w_* = 0 \quad \Rightarrow \quad (H-\lambda_* I)w_* = 0$$

because $H-\lambda_* I$ is real symmetric. Therefore $\lambda_* \in \mathrm{eig}(H)$, which yields $\lambda_* = \lambda_{\min}(H)$. Note that x_* is unique and w_* can be chosen arbitrarily in the eigenspace of H corresponding with eigenvalue $\lambda_{\min}(H)$, so w_* is not unique. Therefore, $y_* = x_* + \sqrt{\gamma^2 - \|x_*\|^2}(w_*/\|w_*\|)$ is unique if and only if $\|x_*\| = \gamma$. □

Remark 2.1. In [10], the authors investigate the relationship between the problems

$$\text{pLG:} \quad (H-\lambda I)y = -g_0, \ \ \|y\| = \gamma, \tag{2.32}$$

$$\text{pQEP:} \quad (H-\lambda I)^2 w = \gamma^{-2}g_0g_0^{\mathrm{T}}w, \ \ \lambda \in \mathbb{R}, \ \ w \neq 0. \tag{2.33}$$

They differ from pLGopt and pQEPmin without taking the min over λ. The following results were obtained in [10]:

1. If (λ,y) is a solution of pLG (2.32), then there exists w such that (λ,w) is a solution of pQEP (2.33).

2. Suppose that (λ,w) is a solution of pQEP (2.33).
   - If $\lambda \notin \mathrm{eig}(H)$, then there exists y such that (λ,y) is a solution of pLG (2.32).
   - If $\lambda \in \mathrm{eig}(H)$, then there exists y such that (λ,y) is a solution of pLG (2.32) if and only if $\|(H-\lambda I)^{\dagger}g_0\| \leqslant \gamma$.

Consequently, these results provide no guarantee that for any solution (λ,w) of pQEP (2.33), there exists a corresponding solution (λ,y) of pLG (2.32). Nonetheless, the authors stated without proof that for the solution (λ_*,w_*) of pQEP (2.33) with λ_* being the smallest eigenvalue of pQEP (2.33), there does exist a solution (λ_*,y_*) of pLGopt (2.23), a conclusion that does not seem straightforward. In Theorem 2.5 we rigorously proved that for any minimizer (λ_*,w_*) of pQEPmin (2.26), there exists y_* such that (λ_*,y_*) is a minimizer of pLGopt (2.23).

Next we will establish an important result in Theorem 2.7 below that says the leftmost eigenvalue of QEP (2.26b) is real. We begin by establishing a close relationship between the zeros of the secular function $\chi(\lambda)$ in (2.29) and the eigenvalues of QEP (2.26b), and then using the relation to expose an eigenvalue distribution property of QEP (2.26b).

Lemma 2.7. *Suppose $\lambda\notin\operatorname{eig}(H)$, λ (possibly complex) is an eigenvalue of* QEP (2.26b) *if and only if $\chi(\lambda)=0$, where $\chi(\lambda)$ is defined in* (2.29).

Proof. Let $\chi(\lambda)=0$ and $\lambda\notin\operatorname{eig}(H)$. Define $z=(H-\lambda I)^{-2}g_0$. Then we have $(H-\lambda I)^2z=g_0$ and

$$g_0^{\mathrm{T}}z=\sum_{i=1}^{\ell}\frac{\xi_i^2}{(\theta_i-\lambda)^2}=\gamma^2 \quad\text{and thus}\quad (H-\lambda I)^2z=g_0=\gamma^{-2}g_0g_0^{\mathrm{T}}z,$$

i.e., (λ,z) is an eigenpair of QEP (2.26b).

On the other hand, suppose λ is an eigenvalue of QEP (2.26b) and $\lambda\notin\operatorname{eig}(H)$. Premultiply (2.26b) by $g_0^{\mathrm{T}}(H-\lambda I)^{-2}$ to get

$$g_0^{\mathrm{T}}z=\gamma^{-2}g_0^{\mathrm{T}}(H-\lambda I)^{-2}g_0g_0^{\mathrm{T}}z. \tag{2.34}$$

We claim that $g_0^{\mathrm{T}}z\neq 0$. Otherwise, $(H-\lambda I)^2z=0$ by (2.26b), which implies $(H-\lambda I)z=0$, i.e., $\lambda\in\operatorname{eig}(H)$, a contradiction. So $g_0^{\mathrm{T}}z\neq 0$ and thus it follows from (2.34) that

$$\gamma^{-2}g_0^{\mathrm{T}}(H-\lambda I)^{-2}g_0=1,$$

i.e., λ is a zero of $\chi(\lambda)$, as was to be shown. □

Lemma 2.8. QEP (2.26b) *has no eigenvalue $\lambda=\alpha+\mathtt{i}\beta$ with $\alpha<\theta_{j_0}$ and $\beta\neq 0$, where $\alpha,\beta\in\mathbb{R}$, $\mathtt{i}$ is the imaginary unit, and j_0 is defined in* (2.30).

Proof. Suppose, to the contrary, that QEP (2.26b) has an eigenvalue $\lambda=\alpha+\mathtt{i}\beta$ with $\alpha<\theta_{j_0}$ and $\beta\neq 0$. Evidently $\lambda=\alpha+\mathtt{i}\beta\notin\operatorname{eig}(H)$ because all eigenvalues of H are real. By

Lemma 2.7, $\alpha+\mathtt{i}\beta$ must be a zero of the secular function $\chi(\lambda)$ in (2.29), i.e.,

$$
\begin{aligned}
0=\chi(\alpha+\mathtt{i}\beta) &= \sum_{i=1}^{\ell}\frac{\xi_i^2}{(\alpha-\theta_i+\mathtt{i}\beta)^2}-\gamma^2 \\
&= \sum_{i=1}^{\ell}\frac{\xi_i^2}{(\alpha-\theta_i)^2-\beta^2+2\mathtt{i}(\alpha-\theta_i)\beta}-\gamma^2 \\
&= \sum_{i=1}^{\ell}\frac{\xi_i^2[(\alpha-\theta_i)^2-\beta^2-2\mathtt{i}(\alpha-\theta_i)\beta]}{[(\alpha-\theta_i)^2-\beta^2]^2+4\beta^2(\alpha-\theta_i)^2}-\gamma^2.
\end{aligned}
$$

In particular, the imaginary part of $\chi(\alpha+\mathtt{i}\beta)$ is zero, i.e.,

$$
\sum_{i=1}^{\ell}\frac{-2(\alpha-\theta_i)\beta\xi_i^2}{[(\alpha-\theta_i)^2-\beta^2]^2+4\beta^2(\alpha-\theta_i)^2}=\beta\left(\sum_{i=j_0}^{\ell}\frac{-2(\alpha-\theta_i)\xi_i^2}{[(\alpha-\theta_i)^2-\beta^2]^2+4\beta^2(\alpha-\theta_i)^2}\right)=0. \tag{2.35}
$$

Since $\alpha<\theta_i$ for all $i\geqslant j_0$, $\xi_{j_0}^2>0$ and $\xi_i^2\geqslant 0$ for all $i>j_0$, we know

$$
\sum_{i=j_0}^{\ell}\frac{-2(\alpha-\theta_i)\xi_i^2}{[(\alpha-\theta_i)^2-\beta^2]^2+4\beta^2(\alpha-\theta_i)^2}>0.
$$

Therefore, by (2.35), we conclude $\beta=0$, a contradiction. □

Lemma 2.9. QEP (2.26b) *has an eigenvalue* $\widetilde{\lambda}<\theta_{j_0}$ *(necessarily* $\widetilde{\lambda}\in\mathbb{R}$*), where* j_0 *is defined in* (2.30).

Proof. There are two possible cases:

- Case $\theta_{j_0}=\theta_1$: Without loss of generality, let $\xi_1\neq 0$. Since $\chi(\lambda)$ is continuous and strictly increasing in $(-\infty,\theta_1)$, and

$$
\lim_{\lambda\to-\infty}\chi(\lambda)=-\gamma^2<0,\quad \lim_{\lambda\to\theta_1^-}\chi(\lambda)\geqslant\lim_{\lambda\to\theta_1^-}\frac{\xi_1^2}{(\lambda-\theta_1)^2}-\gamma^2=+\infty>0,
$$

  there exists a zero $\widetilde{\lambda}\in(-\infty,\theta_1)$ of $\chi(\lambda)$. Evidently $\widetilde{\lambda}\notin\operatorname{eig}(H)$, and then by Lemma 2.7, $\widetilde{\lambda}$ must be an eigenvalue of QEP (2.26b).

- Case $\theta_{j_0}>\theta_1$: Let $\widetilde{\lambda}=\theta_1$ and $z=y_1$. We have $(H-\widetilde{\lambda}I)^2z=(H-\widetilde{\lambda}I)^2y_1=0$. Furthermore, $g_0^{\mathrm{T}}z=g_0^{\mathrm{T}}y_1=\xi_1=0$. Therefore $(\widetilde{\lambda},z)$ satisfies (2.26b), implying $\widetilde{\lambda}$ is an eigenvalue of QEP (2.26b) and $\tilde{\lambda}=\theta_1<\theta_{j_0}$.

The proof is completed. □

With the three lemmas above, now we are ready to prove our main result on the leftmost eigenvalue of QEP (2.26b).

Theorem 2.7. *The leftmost eigenvalue, by which we mean the one with the smallest real part, of* QEP (2.26b) *is real. As a consequence, the optimal value of* pQEPmin (2.26) λ_* *is the leftmost eigenvalue of* QEP (2.26b).

Proof. Let $\lambda_*=\alpha_*+\mathtt{i}\beta_*$ be the leftmost eigenvalue. By Lemma 2.9, QEP (2.26b) has a real eigenvalue $\widetilde{\lambda}$ with $\widetilde{\lambda}<\theta_{j_0}$. Hence $\alpha_*\leqslant\widetilde{\lambda}<\theta_{j_0}$, which together with Lemma 2.8 tell us that $\beta_*=0$ and thus $\lambda_*\in\mathbb{R}$. □

Remark 2.2. In [37], the authors stated without proof that the rightmost eigenvalue of the QEP

$$\left[(W+\lambda I)^2-\delta^{-2}hh^{\mathrm{T}}\right]x=0 \tag{2.36}$$

is real and positive, where W is a real symmetric matrix, h is a vector, and $\delta>0$ is a scalar. It was pointed out in [20] that the rightmost eigenvalue of (2.36) may not always be positive and the authors proved in [20, Theorem 4.1] that the largest real eigenvalue of (2.36) is the rightmost eigenvalue. The authors applied a maximin principle for nonlinear eigenproblems for the proof. In Theorem 2.7 we have proved the leftmost eigenvalue λ_* of (2.26b) is real, i.e., there is no complex eigenvalue of QEP (2.26b) with real part equal to λ_* and nonzero complex part. This result cannot be obtained by the approach used in [20].

2.7 The equivalence of LGopt and QEPmin

Theorem 2.5 says that pLGopt (2.23) and pQEPmin (2.26) are equivalent. Previously in Lemma 2.2, we showed that LGopt (2.13) and QEPmin (2.18) are also equivalent under the assumptions stated there. Our goal in this subsection is to have the assumptions of Lemma 2.2 removed.

For convenience, we restate LGopt (2.13) and QEPmin (2.18) as follows:

$$\text{LGopt:}\quad\left\{\begin{array}{ll}\min\lambda & \text{(2.13a)}\\ \quad\text{s.t. } (PAP-\lambda I)u=-b_0, & \text{(2.13b)}\\ \qquad\|u\|=\gamma, & \text{(2.13c)}\\ \qquad u\in\mathcal{N}(C^{\mathrm{T}}); & \text{(2.13d)}\end{array}\right.$$

and

$$\text{QEPmin:}\quad\left\{\begin{array}{ll}\min\lambda & \text{(2.18a)}\\ \quad\text{s.t. } (PAP-\lambda I)^2z=\gamma^{-2}b_0b_0^{\mathrm{T}}z, & \text{(2.18b)}\\ \qquad\lambda\in\mathbb{R},\ 0\neq z\in\mathcal{N}(C^{\mathrm{T}}). & \text{(2.18c)}\end{array}\right.$$

Recall S_1 and S_2 as defined in (2.19) and H and g as defined in (2.21). Before stating our main result in this subsection, we need two lemmas. The first one is about an eigen-relationship between PAP and H and the second one is on the relationships among $PAP-\lambda I$, $H-\lambda I$, $(PAP-\lambda I)^\dagger$ and $(H-\lambda I)^\dagger$.

Lemma 2.10. *(λ,s) is an eigenpair of H if and only if (λ,S_1s) is an eigenpair of PAP with $S_1s\in\mathcal{N}(C^{\mathrm{T}})$.*

Proof. This is a consequence of the decomposition (2.22a). □

Lemma 2.11. *For any $\lambda\in\mathbb{R}$, $(PAP-\lambda I)S_1=S_1(H-\lambda I)$ and $(PAP-\lambda I)^\dagger S_1=S_1(H-\lambda I)^\dagger$.*

Proof. Let $H=Y\Theta Y^{\mathrm{T}}$ be the eigen-decomposition of H, where $Y\in\mathbb{R}^{(n-m)\times(n-m)}$ is orthogonal and Θ is a diagonal matrix. Then the eigen-decomposition of PAP is given by

$$PAP=[S_1\ S_2]\begin{bmatrix} Y & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Theta & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} Y^{\mathrm{T}} & 0\\ 0 & I\end{bmatrix}[S_1\ S_2]^{\mathrm{T}}. \tag{2.37}$$

Therefore $(PAP-\lambda I)S_1=S_1Y(\Theta-\lambda I)Y^{\mathrm{T}}=S_1(H-\lambda I)$. On the other hand, for $\lambda\neq 0$,

$$(PAP-\lambda I)^\dagger=[S_1\ S_2]\begin{bmatrix} Y & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}(\Theta-\lambda I)^\dagger & 0\\ 0 & -\frac{1}{\lambda}I\end{bmatrix}\begin{bmatrix} Y^{\mathrm{T}} & 0\\ 0 & I\end{bmatrix}[S_1\ S_2]^{\mathrm{T}},$$

and for $\lambda=0$,

$$(PAP)^\dagger=[S_1\ S_2]\begin{bmatrix} Y & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Theta^\dagger & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} Y^{\mathrm{T}} & 0\\ 0 & I\end{bmatrix}[S_1\ S_2]^{\mathrm{T}}.$$

Hence $(PAP-\lambda I)^\dagger S_1=S_1Y(\Theta-\lambda I)^\dagger Y^{\mathrm{T}}=S_1(H-\lambda I)^\dagger$, as was to be shown. □

Now we are ready to state the main result of the subsection.

Theorem 2.8 (The equivalence of LGopt (2.13) **and QEPmin** (2.18)**).**

(1) *Let (λ_*,u_*) be a minimizer of* LGopt (2.13). *Then there exists z_* such that (λ_*,z_*) is a minimizer of* QEPmin (2.18). *Specifically,*

$$z_*=\begin{cases}(PAP-\lambda_*I)^\dagger u_*, & \text{if } \lambda_*\notin\operatorname{eig}(PAP) \text{ or } \lambda_*\in\operatorname{eig}(PAP) \text{ but there is no corre-}\\ & \text{sponding eigenvector entirely in } \mathcal{N}(C^{\mathrm{T}}),\\ s, & \text{if } \lambda_*\in\operatorname{eig}(PAP) \text{ and there is a corresponding eigenvector}\\ & s\in\mathcal{N}(C^{\mathrm{T}}).\end{cases}$$

(2) *Let (λ_*,z_*) be a minimizer of* QEPmin (2.18). *Then there exists $u_*\in\mathbb{R}^n$ such that (λ_*,u_*) is a minimizer of* LGopt (2.13). *Specifically,*

$$u_*=\begin{cases}-(\gamma^2/b_0^{\mathrm{T}}z_*)(PAP-\lambda_*I)z_*, & \text{if } b_0^{\mathrm{T}}z_*\neq 0,\\ x_*+\sqrt{\gamma^2-\|x_*\|^2}(z_*/\|z_*\|), & \text{if } b_0^{\mathrm{T}}z_*=0,\end{cases}$$

where $x_*=-(PAP-\lambda_*I)^\dagger b_0$ in the case $b_0^{\mathrm{T}}z_*=0$ and it is guaranteed that $\|x_*\|\leqslant\gamma$.

Proof. We prove item (1) first. By Theorem 2.3, (λ_*,y_*) with $y_*=S_1^{\mathrm{T}}u_*$ is a minimizer of pLGopt (2.23). We have two cases to consider.

(a) If $\lambda_*\notin\operatorname{eig}(PAP)$ or $\lambda_*\in\operatorname{eig}(PAP)$ but there is no corresponding eigenvector $s\in\mathcal{N}(C^{\mathrm{T}})$, then $\lambda_*\notin\operatorname{eig}(H)$ by Lemma 2.10. Using Theorem 2.5, we conclude that (λ_*,w_*) with

$$w_*=(H-\lambda_*I)^{-1}y_*=(H-\lambda_*I)^{\dagger}y_*$$

is a minimizer of pQEPmin (2.26). Now use Theorem 2.4 to conclude that (λ_*,z_*) with $z_*=S_1(H-\lambda_*I)^{\dagger}y_*$ is a minimizer of QEPmin (2.18). By Lemma 2.11,

$$z_*=S_1(H-\lambda_*I)^{\dagger}y_*=(PAP-\lambda_*I)^{\dagger}S_1y_*=(PAP-\lambda_*I)^{\dagger}u_*.$$

(b) Suppose that $\lambda_*\in\operatorname{eig}(PAP)$ and there is a corresponding eigenvector $s\in\mathcal{N}(C^{\mathrm{T}})$. Then $s=S_1r$ for some $0\neq r\in\mathbb{R}^{n-m}$. By Lemma 2.10, r is an eigenvector of H corresponding to the eigenvalue λ_*. Use Theorem 2.5 to conclude that (λ_*,w_*) with $w_*=r$ is a minimizer of pQEPmin (2.26), which in turn, by Theorem 2.4, yields that (λ_*,z_*) with $z_*=s=S_1r$ is a minimizer of QEPmin (2.18).

Next we consider item (2). By Theorem 2.4, (λ_*,w_*) with $w_*=S_1^{\mathrm{T}}z_*$ is a minimizer of pQEPmin (2.26). Since $b_0,z_*\in\mathcal{N}(C^{\mathrm{T}})$, we have $z_*=S_1w_*$ and $b_0^{\mathrm{T}}z_*=g_0^{\mathrm{T}}S_1^{\mathrm{T}}S_1w_*=g_0^{\mathrm{T}}w_*$.

- Case $b_0^{\mathrm{T}}z_*\neq0$: We have $g_0^{\mathrm{T}}w_*\neq0$. By Theorem 2.5, (λ_*,y_*) with $y_*=-(\gamma^2/g_0^{\mathrm{T}}w_*)(H-\lambda_*I)w_*$ solves pLGopt (2.23). By Theorem 2.3, (λ_*,u_*) with $u_*=-(\gamma^2/g_0^{\mathrm{T}}w_*)S_1(H-\lambda_*I)w_*$ solves LGopt (2.13). Furthermore, by Lemma 2.11, $(PAP-\lambda_*I)z_*=(PAP-\lambda_*I)S_1w_*=S_1(H-\lambda_*I)w_*$. Therefore $u_*=-(\gamma^2/g_0^{\mathrm{T}}w_*)S_1(H-\lambda_*I)w_*=-(\gamma^2/b_0^{\mathrm{T}}z_*)(PAP-\lambda_*I)z_*$.

- Case $b_0^{\mathrm{T}}z_*=0$: We have $g_0^{\mathrm{T}}w_*=0$ and z_* is an eigenvector of PAP corresponding to its eigenvalue λ_*. By Lemma 2.10, $y_*=S_1^{\mathrm{T}}z_*$ is an eigenvector of H corresponding to its eigenvalue λ_*. Let $s=-(H-\lambda_*I)^{\dagger}g$, according to Theorem 2.5, $\|s\|\leqslant\gamma$ and (λ_*,y_*) with $y_*=s+\sqrt{\gamma^2-\|s\|^2}(w_*/\|w_*\|)$ solves pLGopt (2.23). By Theorem 2.4, (λ_*,u_*) with $u_*=S_1y_*$ is a minimizer of LGopt (2.13). Now set

$$x_*=S_1s=-S_1(H-\lambda_*I)^{\dagger}g=-(PAP-\lambda_*I)^{\dagger}b_0,$$

and thus

$$u_*=S_1y_*=S_1s+\sqrt{\gamma^2-\|S_1s\|^2}\,\frac{S_1w_*}{\|S_1w_*\|}=x_*+\sqrt{\gamma^2-\|x_*\|^2}\,\frac{z_*}{\|z_*\|},$$

as expected.

This completes the proof. □

We note that proving the equivalence between LGopt (2.13) and QEPmin (2.18) is of theoretical interest. The proof in [10] is incomplete as discussed in Remark 2.1.

Returning to the original CRQopt (1.1), we observe that if (λ_*,u_*) solves LGopt (2.13), then n_0+u_* solves CRQopt (1.1). Therefore immediately we obtain the following theorem.

Theorem 2.9. *Suppose* (λ_*,z_*) *is a minimizer of* QEPmin (2.18). *Then a minimizer* v_* *of* CRQopt (1.1) *is given by*

$$v_*=\begin{cases} n_0-(\gamma^2/b_0^{\mathrm T}z_*)(PAP-\lambda_*I)z_*, & \textit{if } b_0^{\mathrm T}z_*\neq 0,\\ n_0+x_*+\sqrt{\gamma^2-\|x_*\|^2}(z_*/\|z_*\|), & \textit{if } b_0^{\mathrm T}z_*=0,\end{cases}$$

where $x_*=-(PAP-\lambda_*I)^\dagger b_0$ *in the case of* $b_0^{\mathrm T}z_*=0$ *and it is guaranteed that* $\|x_*\|\leqslant\gamma$.

What the next theorem says is that solving QEPmin (2.18) is equivalent to calculating the leftmost eigenvalue of QEP (2.18b) among those having eigenvectors[§] in $\mathcal{N}(C^{\mathrm T})$. This result paves the way for the use of a Krylov subspace method to calculate the minimizer of QEPmin (2.18) in Section 3 ahead.

Theorem 2.10. *If* (λ_*,z_*) *is a minimizer of* QEPmin (2.18), *then* λ_* *is the leftmost eigenvalue of* QEP (2.18b) *among those having eigenvectors in* $\mathcal{N}(C^{\mathrm T})$.

Proof. Following the argument in the proof of Theorem 2.4, we find that the set of eigenvalues of QEP (2.18b) that have eigenvectors in $\in\mathcal{N}(C^{\mathrm T})$ and the set of eigenvalues of QEP (2.26b) are the same. The conclusion is an immediate consequence of Theorems 2.4 and 2.7. □

2.8 Summary

Starting with CRQopt (1.1), we have introduced five equivalent optimization problems. Fig. 1 summarizes the relationships of these problems. The edge "⟷" in Fig. 1 connecting two optimization problems indicates that we have an equivalent relationship in the previous subsections. We note that CRQopt (1.1) and CQopt (2.5) share the same minimizers v_*, while correspondingly the minimizer for LGopt (2.13) is $u_*=Pv_*$. Slightly more efforts are needed to describe corresponding minimizers for other equivalent optimization problems. The optimal values for the objective functions of LGopt (2.13), pLGopt (2.23), QEPmin (2.18), and pQEPmin (2.26) are all the same. The proof of Theorem 2.8 relies on Theorems 2.3, 2.4 and 2.5.

2.9 Easy and hard cases

Motivated by the treatments of the trust-region subproblem [27,43], QEPmin (2.18) can be classified into two categories, namely *easy* case and *hard* case, defined as follows.

[§] This does not exclude the possibility that they may have eigenvectors not in $\mathcal{N}(C^{\mathrm T})$.

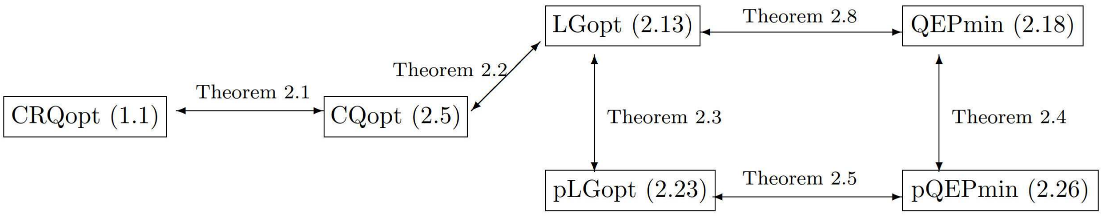

Figure 1: Equivalence of optimization problems.

Definition 2.1. *QEPmin* (2.18) *is in the* hard *case if it has a minimizer* (λ_*,z_*) *with* $b_0^{\mathrm{T}}z_*=0$. *Otherwise, QEPmin* (2.18) *is in the* easy *case. Furthermore, any one of the equivalent optimization problems as shown in Fig.* 1 *is said to be in the* hard *or* easy *case if the corresponding* QEPmin *is.*

The notion of hardness and easiness has its historical reason in dealing with the trust-region subproblem. The hard case is not really hard as its name suggests when it comes to numerical computation. It is just a degenerate and rare case that needs special attention. The easy case is a generic one. Consider the hard case, let $\mathcal{V}$ be the maximal eigenspace of PAP corresponding to eigenvalue λ_*, then $b_0\perp\mathcal{V}$ by Theorem 2.11. This creates difficulties to our later Lanczos method to solve QEPmin (2.18) in that the Krylov subspace $\mathcal{K}_k(PAP,b_0)\subset\mathcal{V}^{\perp}$ for any k. So in theory there is no vector in $\mathcal{K}_k(PAP,b_0)$ can approximate any eigenvector $z\in\mathcal{V}$ well.

In Theorems 2.11 and 2.12 below, we present a number of characterizations about the *hard* case.

Lemma 2.12. QEPmin (2.18) *is in the* hard *case if and only if* pQEPmin (2.26) *has a minimizer* (λ_*,w_*) *satisfying* $g_0^{\mathrm{T}}w_*=0$.

Proof. To see this, we let (λ_*,z_*) be a minimizer QEPmin (2.18) satisfying $b_0^{\mathrm{T}}z_*=0$. By Theorem 2.4, we know that z_* and w_* are related by $z_*=S_1w_*$. Since also $b_0=S_1g_0$, $b_0^{\mathrm{T}}z_*=g_0^{\mathrm{T}}w_*$. □

Theorem 2.11. *Suppose that* QEPmin (2.18) *is in the* hard *case, and let* (λ_*,z_*) *be a minimizer such that* $b_0^{\mathrm{T}}z_*=0$. *Then we have the following statements:*

(1) $\lambda_*=\lambda_{\min}(H)$, *the smallest eigenvalue of* H;

(2) $g_0\perp\mathcal{U}$, *where* $\mathcal{U}$ *is the eigenspace of* H *associated with its eigenvalue* $\lambda_{\min}(H)$;

(3) $b_0\perp\mathcal{V}$, *where* $\mathcal{V}$ *is the eigenspace of* PAP *associated with its eigenvalue* $\lambda_{\min}(H)\in\operatorname{eig}(PAP)$.

Proof. By Lemma 2.12, pQEPmin (2.26) has a minimizer (λ_*,w_*) satisfying $g_0^{\mathrm{T}}w_*=0$. Theorem 2.6 immediately leads to item (1). Item (2) is a corollary of Lemma 2.4.

For item (3), it follows from Lemma 2.3 that if $\lambda_{\min}(H)\neq 0$, then $\mathcal{V}=S_1\mathcal{U}$. Since $b_0=S_1g_0$ and $g_0\perp\mathcal{U}$ by item (2), we conclude that $b_0\perp S_1\mathcal{U}$. If, however, $\lambda_{\min}(H)=0$, then $\mathcal{V}=S_1\mathcal{U}+\mathcal{R}(S_2)$. Since again $g_0\perp\mathcal{U}$ by item (2) and also $b_0\perp\mathcal{R}(S_2)$, we still have $b_0\perp\mathcal{V}$. □

Theorem 2.12. QEPmin (2.18) *is in the* hard *case if and only if*

$$g_0 \perp \mathcal{U} \quad \text{and} \quad \|[H-\lambda_{\min}(H)I]^{\dagger}g_0\|_2 \leqslant \gamma, \tag{2.38}$$

where $\mathcal{U}$ is as defined in Theorem 2.11.

Proof. If QEPmin (2.18) is in the *hard* case, then its optimal value (which is also the one of LGopt (2.13)) $\lambda_* = \lambda_{\min}(H)$. This can only happen when (2.38) holds. On the other hand, if (2.38) holds, then $\lambda_* = \lambda_{\min}(H)$ by Lemma 2.4. By Theorem 2.5, pQEPmin (2.26) has a minimizer (λ_*, w_*), where $Hw_* = \lambda_* w_*$. Thus $g_0^{\mathrm{T}} w_* = 0$ because $g_0 \perp \mathcal{U}$ and $w_* \in \mathcal{U}$. Hence QEPmin (2.18) is in the *hard* case by Lemma 2.12. □

When QEPmin (2.18) is in the easy case, the situation is much simpler to characterize.

Theorem 2.13. CRQopt (1.1) *has a unique minimizer when* QEPmin (2.18) *is in the easy case.*

Proof. Suppose that QEPmin (2.18) is in the easy case. By Definition 2.1, all minimizers (λ_*, w_*) of pQEPmin (2.26) satisfy $g_0^{\mathrm{T}} w_* \neq 0$. Theorem 2.6 guarantees that pLGopt (2.23) has a unique minimizer. Consequently, the minimizer of LGopt (2.13) is unique by Theorem 2.3 and so is the minimizer of CRQopt (1.1). □

We use the remaining part of this subsection to explain how CRQopt (1.1) and the well-known trust-region subproblem (TRS) are related. We have already proved in Theorem 2.1 that CRQopt (1.1) is equivalent to CQopt (2.5). Set $u = Pv$. Solving CQopt (2.5) is equivalent to solving

$$\begin{cases} \min\ u^{\mathrm{T}} PAPu + 2u^{\mathrm{T}} b_0, & (2.39\text{a}) \\ \quad \text{s.t.}\ \|u\| = \gamma, & (2.39\text{b}) \\ \qquad u \in \mathcal{N}(C^{\mathrm{T}}). & (2.39\text{c}) \end{cases}$$

Let H and g_0 be defined in (2.21) and S_1 be defined in (2.19). Then u is a minimizer of optimization problem (2.39) if and only if $y = S_1^{\mathrm{T}} u$ is a minimizer of the following equality constrained optimization problem

$$\begin{cases} \min\ y^{\mathrm{T}} Hy + 2y^{\mathrm{T}} g_0, & (2.40\text{a}) \\ \quad \text{s.t.}\ \|y\| = \gamma. & (2.40\text{b}) \end{cases}$$

The Lagrange equations for (2.40) is exactly the same as pLGopt (2.23). The problem (2.40) is similar to TRS

$$\begin{cases} \min\ y^{\mathrm{T}} Hy + 2y^{\mathrm{T}} g_0, & (2.41\text{a}) \\ \quad \text{s.t.}\ \|y\| \leqslant \gamma, & (2.41\text{b}) \end{cases}$$

except that its constraint is an equality instead of an inequality. When H is not positive semi-definite, solution of (2.40) and TRS (2.41) are exactly the same. But when H is positive semi-definite and $g_0 \perp \mathcal{N}(H)$, we need to check whether $\|H^{\dagger} g_0\| < \gamma$. If so, $-H^{\dagger} g_0$, instead of the minimizer of (2.40), is the minimizer of TRS (2.41). If, however, $\|H^{\dagger}g_0\| \geqslant \gamma$, then the minimizer of TRS (2.41) is the same as that of (2.40).

Lemma 2.1 in [17] shows that y is the solution of (2.40) if and only if there exists $\widehat{\lambda} \in \mathbb{R}$ such that $(\widehat{\lambda}, y)$ satisfies the constraints of pLGopt (2.23) and $H - \widehat{\lambda} I$ is positive semi-definite. According to Lemma 2.4, the optimal value of pLGopt (2.23) satisfies $\lambda_* \leqslant \lambda_{\min}(H)$, which indicates that $H - \lambda_* I$ is positive semi-definite. Therefore, solving the equality constrained problem (2.40) is equivalent to solving pLGopt (2.23).

As we have mentioned, the terms "*easy*" and "*hard*" were adopted from the treatments of the trust-region subproblem [27,43], where the term "easy" means the associated case is easy to explain, not implying the case is easy to solve, however. A more detailed connection with TRS (2.41) is as follows.

1. In the easy case of QEPmin (2.18), $b_0^{\mathrm{T}} z_* \neq 0$ for all minimizers (λ_*, z_*). By Theorem 2.4, $z_* = S_1 w_*$ for some $w_* \in \mathbb{R}^{n-m}$ and thus $g_0^{\mathrm{T}} w_* = b_0^{\mathrm{T}} S_1 w_* = b_0^{\mathrm{T}} z_* \neq 0$. By Theorem 2.6, $\lambda_* < \lambda_{\min}(H)$, and thus (λ_*, y_*) with $y_* = -(H - \lambda_* I)^{-1} g_0$ is the unique minimizer of pLGopt (2.23). Hence y_* is the unique minimizer of (2.40), which is related to the easy case of TRS (2.41).

2. In the hard case of QEPmin (2.18), there exists a minimizer (λ_*, z_*) such that $b_0^{\mathrm{T}} z_* = 0$. Again by Theorem 2.4, $z_* = S_1 w_*$ for some $w_* \in \mathbb{R}^{n-m}$ and $g_0^{\mathrm{T}} w_* = 0$. By Theorem 2.5, a minimizer of pLGopt (2.23) is given by

$$(\lambda_*, y_*) \quad \text{and} \quad y_* = x_* + \sqrt{\gamma^2 - \|x_*\|^2} \, \frac{w_*}{\|w_*\|},$$

   where $x_* = -(H - \lambda_* I)^{\dagger} g_0$ and it is guaranteed that $\|x_*\| \leqslant \gamma$. Therefore, in general, a minimizer of (2.40) can be expressed by $y_* = x_* + \sqrt{\gamma^2 - \|x_*\|^2}(w_*/\|w_*\|)$, which is related to the hard case of TRS (2.41).

It is known that the generalized Lanczos method does not work for TRS (2.41) in the hard case [43, Theorem 4.6]. A restarting strategy was proposed to overcome the difficulty, but it was commented that the strategy is computationally expensive for large scale problems [16, Theorem 5.8].

In the next section, we present that the Lanczos algorithms for CRQopt (1.1), which resemble the generalized Lanczos method for TRS and are suitable for handling the easy case. However, with some additional effort, the hard case can be detected.

3 Lanczos algorithm

As was shown in Section 2, solving CRQopt (1.1) is equivalent to solving LGopt (2.13) or QEPmin (2.18). In this section we present algorithms to solve CRQopt (1.1) by solving LGopt (2.13) and QEPmin (2.18). We first review the Lanczos procedure in Section 3.1, and then we apply the procedure to reduce LGopt (2.13) and QEPmin (2.18), and finally solve the reduced LGopt and QEPmin to yield approximations to the minimizer of the original CRQopt (1.1). In addition, we prove the finite step stopping property of the proposed algorithms and comment on how to detect the hard case.

3.1 Lanczos process

We review the standard symmetric Lanczos process [4,13,30,34]. Given a real symmetric matrix $M\in\mathbb{R}^{n\times n}$ and a starting vector $r_0\in\mathbb{R}^n$, the Lanczos process partially computes the decomposition $MQ=QT$, where $T\in\mathbb{R}^{n\times n}$ is symmetric and tridiagonal, $Q\in\mathbb{R}^{n\times n}$ is orthogonal and the first column of Q is parallel to r_0.

Specifically, let $Q=[q_1,q_2,\cdots,q_n]$ and denote by α_i for $1\leqslant i\leqslant n$ the diagonal entries of T, and by β_i for $2\leqslant i\leqslant n$ the sub-diagonal and super-diagonal entries of T. The Lanczos process goes as follows: set $q_1=r_0/\|r_0\|$, and equate the first column of both sides of the equation $MQ=QT$ to get

$$Mq_1=q_1\alpha_1+q_2\beta_2. \tag{3.1}$$

Pre-multiply both sides of the equation (3.1) by $q_1^{\rm T}$ to get $\alpha_1=q_1^{\rm T}Mq_1$, and then let

$$\widehat{q}_2=Mq_1-q_1\alpha_1,\quad \beta_2=\|\widehat{q}_2\|.$$

If $\beta_2>0$, set $q_2=\widehat{q}_2/\beta_2$; otherwise the process breaks down. In general for $j\geqslant 2$, equating the jth column of both sides of the equation $MQ=QT$ leads to

$$Mq_j=q_{j-1}\beta_j+q_j\alpha_j+q_{j+1}\beta_{j+1}. \tag{3.2}$$

Up to this point, q_i for $1\leqslant i\leqslant j$, α_i for $1\leqslant i\leqslant j-1$, and β_i for $2\leqslant i\leqslant j$ have already been determined. Pre-multiply both sides of Eq. (3.2) by $q_j^{\rm T}$ to get $\alpha_j=q_j^{\rm T}Mq_j$, and then let

$$\widehat{q}_{j+1}=Mq_j-q_{j-1}\beta_j-q_j\alpha_j,\quad \beta_{j+1}=\|\widehat{q}_{j+1}\|.$$

Now if $\beta_{j+1}>0$, we set $q_{j+1}=\widehat{q}_{j+1}/\beta_{j+1}$; otherwise the process breaks down. The process can be compactly expressed by[¶]

$$MQ_k=Q_kT_k+\beta_{k+1}q_{k+1}e_k^{\rm T} \tag{3.3}$$

assuming the process encounters no breakdown for the first k steps, i.e., no $\beta_i=0$ for $2\leqslant i\leqslant k$, where

$$Q_k=[q_1,q_2,\cdots,q_k],\quad T_k=Q_k^{\rm T}MQ_k=\begin{bmatrix}\alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{k-1} & \alpha_{k-1} & \beta_k \\ & & & \beta_k & \alpha_k\end{bmatrix}.$$

[¶] We sacrifice slightly mathematical rigor in writing (3.3) in exchange for simplicity and convenience, since q_{k+1} cannot be determined unless also $\beta_{k+1}>0$.

Furthermore, the column space $\mathcal{R}(Q_k)$ is the same as the kth Krylov subspace

$$\mathcal{K}_k(M,r_0) := \operatorname{span}(r_0, Mr_0, \cdots, M^{k-1}r_0).$$

In the case of a breakdown with $\beta_{k+1}=0$, $MQ_k=Q_kT_k$ and $\mathcal{R}(Q_k)$ is an invariant subspace of M.

3.2 Solving LGopt

In this subsection, we first use (3.3) obtained by the Lanczos process with $M=PAP$ to reduce LGopt (2.13), and then solve the reduced LGopt via an approach based on a secular equation solver.

3.2.1 Dimensional reduction of LGopt

For the dimensional reduction of LGopt (2.13), we restate the Lagrange equations (2.13b) and (2.13b) here

$$(PAP-\lambda I)u=-b_0, \quad \|u\|=\gamma, \quad Pu=u, \tag{3.4}$$

where we include the constraint $Pu=u$ since we are only interested in those vectors $u\in\mathcal{N}(C^{\mathrm{T}})$.

Apply the Lanczos process with $M=PAP$ and the starting vector $r_0=b_0$ to get (3.3) with $M=PAP$. It then follows that for any scalar λ,

$$Q_k^{\mathrm{T}}(PAP-\lambda I)Q_k=T_k-\lambda I \quad \text{and} \quad Q_k^{\mathrm{T}}b_0=\|b_0\|e_1.$$

Consequently, we arrive at the reduced LGopt (2.13)

$$\text{rLGopt:} \quad \begin{cases} \min\ \lambda & (3.5\text{a}) \\ \text{ s.t. } (T_k-\lambda I)x=-\|b_0\|e_1, & (3.5\text{b}) \\ \qquad \|x\|=\gamma. & (3.5\text{c}) \end{cases}$$

A couple of comments are in order for the efficiency of the Lanczos process with $M=PAP$. In the process, we have to calculate matrix-vector products $Mx=P(A(Pq_j))$ efficiently. For that purpose, it suffices for us to be able to calculate the product Pc efficiently for any given $c\in\mathbb{R}^n$. In fact

$$Pc=c-CC^{\dagger}c=c-Cy,$$

where $y=C^{\dagger}c$ is the minimum-norm solution of the least squares problem

$$y=\arg\min_{z\in\mathbb{R}^m}\|Cz-c\|_2, \tag{3.6}$$

which can be computed by using the QR decomposition of $C\in\mathbb{R}^{n\times m}$ or an iterative method such as LSQR [7, 29, 35]. Another cost-saving observation due to [14] is that for the matrix-vector product $Mq_j = P(A(Pq_j))$, the first application of P in Pq_j can be skipped due to the fact that if the initial vector $b_0 \in \mathcal{N}(C^{\mathrm{T}})$, then $Pq_j = q_j$ for all $1 \leqslant j \leqslant k+1$.

We end this subsection by pointing out that rLGopt (3.5) cannot fall into the hard case. The same phenomenon happens to the tridiagonal TRS generated by the generalized Lanczos method [16, Theorem 5.3] as well. Let the eigen-decomposition of T_k be

$$T_k = Y\Theta Y^{\mathrm{T}}, \quad Y^{\mathrm{T}}Y = I_k, \quad \Theta = \operatorname{diag}(\vartheta_1, \vartheta_2, \cdots, \vartheta_k), \tag{3.7}$$

where we suppress the dependency of Y, Θ, and ϑ_j on k for notational convenience. Further, we arrange ϑ_j in nondecreasing order, i.e., $\vartheta_1 \leqslant \vartheta_2 \leqslant \cdots \leqslant \vartheta_k$ and $Y = [y_1, y_2, \cdots, y_k]$.

Theorem 3.1. *Suppose that $\beta_j \neq 0$ for $j = 2,3,\cdots,k$ in the Lanczos process. Let $\mu^{(k)}$ be the optimal value of rLGopt* (3.5)*, then $\mu^{(k)} < \vartheta_1 \equiv \lambda_{\min}(T_k)$, and* rLGopt (3.5) *cannot fall into the hard case.*

Proof. It is well-known that the first components of all eigenvectors y_i of irreducible T_k are nonzero [30, p.140]. In particular, $e_1^{\mathrm{T}} y_1 \neq 0$. Lemma 2.4 immediately leads to $\mu^{(k)} < \vartheta_1$.

Since $\mu^{(k)} < \lambda_{\min}(T_k)$ by Theorem 2.11(1), we conclude that rLGopt cannot fall into the hard case. □

3.2.2 Solving rLGopt

Now we explain how to solve rLGopt (3.5). Suppose that $\beta_j \neq 0$ for $j = 2,3,\cdots,k$, and let the eigen-decomposition of T_k be given by (3.7).

Theorem 3.2. *The optimal value $\mu^{(k)}$ of* rLGopt (3.5) *is the smallest root of the secular function*

$$\widehat{\chi}(\lambda) = \|b_0\|^2 e_1^{\mathrm{T}} (T_k - \lambda I)^{-2} e_1 - \gamma^2 = \sum_{i=1}^{k} \frac{\zeta_i^2}{(\lambda - \vartheta_i)^2} - \gamma^2, \tag{3.8}$$

where $\zeta_i = \|b_0\| e_1^{\mathrm{T}} y_i$ for $i = 1,2,\cdots,k$. Furthermore,

$$(\mu^{(k)}, x^{(k)}) = (\mu^{(k)}, -\|b_0\| (T_k - \mu^{(k)} I)^{-1} e_1) \tag{3.9}$$

is a minimizer of rLGopt (3.5).

Proof. rLGopt (3.5) takes the same form as pLGopt (2.23). By Theorem 3.1, $\mu^{(k)} < \lambda_{\min}(T_k)$. The conclusions of the lemma are now consequences of Lemma 2.4. □

Theorem 3.2 naturally leads to a method for solving rLGopt (3.5) through calculating the smallest root of the secular function $\widehat{\chi}(\lambda)$. Algorithm 1 outlines the method, based on an efficient secular equation solver in Appendix A.

Although Theorem 3.2 assures us that the hard case cannot happen for rLGopt (3.5), cases where $|e_1^{\mathrm{T}} y_1|$ is very tiny are possible. Such a nearly hard case has to be treated with care, a subject of further study.

Algorithm 1 Solving rLGopt (3.5)

```
Input: $T_k \in \mathbb{R}^{k\times k}$, $\|b_0\|$, $\gamma>0$, and tolerance $\epsilon$;
Output: $(\mu^{(k)},x^{(k)})$, approximate minimizer of rLGopt (3.5);
Compute the eigenvalues $\theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_k$ of $T_k$ and the corresponding eigenvectors $y_1,\cdots,y_k$;
$\xi_i \leftarrow \|b_0\| e_1^{\rm T} y_i$ for $i=1,2,\cdots,k$;
$\delta_0 \leftarrow \frac{1}{\gamma}\sqrt{\sum_{i=1}^k \xi_i^2}$, $\alpha^{(0)} \leftarrow \theta_1-\delta_0$, $\beta^{(0)} \leftarrow \theta_1$ and $\eta \leftarrow \gamma^2-\sum_{i=2}^k \frac{\xi_i^2}{([\theta_1-\delta_0]-\theta_i)^2}$;
if $\eta>0$ then $\lambda^{(0)} \leftarrow \theta_1-|\xi_1|/\sqrt{\eta}$ else $\lambda^{(0)} \leftarrow \theta_1-\delta_0/2$;
for $j=0,1,2,\cdots$ do
    $\chi \leftarrow \sum_{i=1}^k \frac{\xi_i^2}{(\lambda^{(j)}-\theta_i)^2}-\gamma^2$;
    if $\chi>0$ then $\alpha^{(j+1)} \leftarrow \alpha^{(j)}$, $\beta^{(j+1)} \leftarrow \lambda^{(j)}$ else $\alpha^{(j+1)} \leftarrow \lambda^{(j)}$, $\beta^{(j+1)} \leftarrow \beta^{(j)}$;
    $a \leftarrow (\lambda^{(j)}-\theta_1)^3 \sum_{i=1}^n \frac{\xi_i^2}{(\lambda^{(j)}-\theta_i)^3}$, $b \leftarrow (\lambda^{(j)}-\theta_1)\sum_{i=1}^n \frac{\xi_i^2}{(\lambda^{(j)}-\theta_i)^3}-\chi$;
    if $b>0$ then
        $\lambda_1 \leftarrow \theta_1-\sqrt{a/b}$;
        if $\lambda_1 \in (\alpha^{(j+1)},\beta^{(j+1)})$ then $\lambda^{(j+1)} \leftarrow \lambda_1$ else $\lambda^{(j+1)} \leftarrow (\alpha^{(j+1)}+\beta^{(j+1)})/2$;
    else
        $\lambda^{(j+1)} \leftarrow (\alpha^{(j+1)}+\beta^{(j+1)})/2$;
    end if
    if $|\lambda^{(j+1)}-\lambda^{(j)}|<\epsilon$ then stop;
end for
return $(\mu^{(k)},x^{(k)})=(\lambda^{(j+1)},-(T_k-\mu^{(k)}I)^{-1}\|b_0\|e_1)$ as a solution of rLGopt (3.5).
```

Remark 3.1. Let us discuss the relationship between solving rLGopt (3.5) and solving TRS by a generalized Lanczos (GLTRS) method proposed in [16]. GLTRS projects a similar problem to (2.39a) and (2.39b) by a Krylov subspace to yield a small-size problem. Ignoring (2.39c) for the moment, we run the Lanczos process with $M=PAP$ and the starting vector be $r_0=b_0$ to generate the orthonormal basis matrix Q_k and the tridiagonal matrix T_k. Since $b_0 \in \mathcal{N}(C^{\rm T})$, it can be verified that $\mathcal{R}(Q_k) \subset \mathcal{N}(C^{\rm T})$, which means that (2.39c) is automatically taken care of. Project (2.39a) and (2.39b) onto the column space of Q_k and we arrive at the following equality constrained optimization problem:

$$\begin{cases} \min\ x^{\rm T}T_k x+2x^{\rm T}\|g_0\|e_1, & (3.10\text{a}) \\ \ \text{s.t. } \|x\|=\gamma. & (3.10\text{b}) \end{cases}$$

Problem (3.10) is similar to the tridiagonal TRS generated by GLTRS except that the constraint here is equality instead of inequality. Solving (3.10) by the method of the Lagrangian multipliers leads to exactly rLGopt (3.5).

3.2.3 Solving LGopt

After computing $(\mu^{(k)},x^{(k)})$, the minimizer of rLGopt (3.5), we deduce an approximate minimizer of LGopt (2.13):

$$(\mu^{(k)},u^{(k)})=(\mu^{(k)},Q_k x^{(k)}). \tag{3.11}$$

It can be verified that

$$\|u^{(k)}\|=\|x^{(k)}\|=\gamma, \quad u^{(k)}\in\mathcal{R}(Q_k)\subset\mathcal{N}(C^{\mathrm{T}}). \tag{3.12}$$

That is the pair in (3.11) satisfies the constraints (2.13c) and (2.13d).

The accuracy of this approximate minimizer $(\mu^{(k)},u^{(k)})$ can be measured by the residual vector

$$r_k^{\mathrm{LGopt}}=(PAP-\mu^{(k)}I)u^{(k)}+b_0. \tag{3.13}$$

For simplicity, we may assume that $(\mu^{(k)},x^{(k)})$ satisfies the constraint of rLGopt (3.5) exactly. In particular $(T_k-\mu^{(k)}I)x^{(k)}=-\|b_0\|e_1$, since it is reasonable to assume that the error in $(\mu^{(k)},u^{(k)})$ as an approximate minimizer of LGopt (2.13) is much larger than the error in $(\mu^{(k)},x^{(k)})$ as the computed minimizer of rLGopt (3.5). Subsequently, we have the following expression for the residual vector r_k^{LGopt}, similar to the one on the generalized Lanczos method for TRS [16].

Proposition 3.1. *Suppose that the approximate minimizer* $(\mu^{(k)},x^{(k)})$ *of* rLGopt (3.5) *satisfies the constraints of* rLGopt (3.5) *exactly. We have*

$$r_k^{\mathrm{LGopt}}=\beta_{k+1}q_{k+1}(e_k^{\mathrm{T}}x^{(k)}). \tag{3.14}$$

Proof. We have by (3.3)

$$\begin{aligned} r_k^{\mathrm{LGopt}}&=(PAP-\mu^{(k)}I)Q_kx^{(k)}+b_0=[Q_k(T_k-\mu^{(k)}I)+\beta_{k+1}q_{k+1}e_k^{\mathrm{T}}]x^{(k)}+b_0\\ &=-Q_k\|b_0\|e_1+\beta_{k+1}q_{k+1}(e_k^{\mathrm{T}}x^{(k)})+b_0=\beta_{k+1}q_{k+1}(e_k^{\mathrm{T}}x^{(k)}), \end{aligned}$$

as was to be shown. □

In deciding if r_k^{LGopt} is sufficiently small, a sensible way is to check some kind of normalized residual. In view of (3.13), a reasonable one is

$$\mathrm{NRes}_k^{\mathrm{LGopt}}:=\frac{\|r_k^{\mathrm{LGopt}}\|}{(\|A\|+|\mu^{(k)}|)\|x^{(k)}\|+\|b_0\|}=\frac{|\beta_{k+1}||e_k^{\mathrm{T}}x^{(k)}|}{(\|A\|+|\mu^{(k)}|)\|x^{(k)}\|+\|b_0\|}=:\delta_k^{\mathrm{LGopt}}. \tag{3.15}$$

The Lanczos process is stopped if $\delta_k^{\mathrm{LGopt}}\leqslant\epsilon$, a prescribed tolerance. In summary, the Lanczos algorithm for solving LGopt (2.13) is given in Algorithm 2.

Algorithm 2 Solving LGopt (2.13)

```
Input: $A\in\mathbb{R}^{n\times n}$, $C\in\mathbb{R}^{n\times m}$, $b_0\in\mathbb{R}^n$, $\gamma>0$, and tolerance $\epsilon$;
Output: $(\mu^{(k)},u^{(k)})$, approximate minimizer of LGopt (2.13);
$\beta_1\leftarrow\|b_0\|$;
if $\beta_1=0$ then stop;
$q_1\leftarrow b_0/\beta_1$, $q_0\leftarrow 0$;
for $k=1,2,\cdots$ do
    $\hat{q}\leftarrow Aq_k$, $\hat{q}\leftarrow P\hat{q}$, $\hat{q}\leftarrow\hat{q}-\beta_k q_{k-1}$;
    $\alpha_k\leftarrow q_k^{\mathrm{T}}\hat{q}$, $\hat{q}\leftarrow\hat{q}-\alpha_k q_k$, $\beta_{k+1}\leftarrow\|\hat{q}\|$;
    compute the minimizer $(\mu^{(k)},x^{(k)})$ of rLGopt (3.5) by Algorithm 1;
    if $\delta_k^{\mathrm{LGopt}}\leqslant\epsilon$ then stop;
    $q_{k+1}\leftarrow\hat{q}/\beta_{k+1}$;
end for
$Q_k=[q_1,q_2,\cdots,q_k]$;
return $(\mu^{(k)},u^{(k)})$ with $u^{(k)}=Q_kx^{(k)}$ as an approximate minimizer of LGopt (2.13).
```

3.3 Solving QEPmin

In this section, we propose a Lanczos algorithm for solving QEPmin (2.18). It follows the same framework as in the previous subsection. First, we reduce QEPmin (2.18) to a smaller problem by projection, and then solve the reduced QEPmin by an eigensolver. One immediate advantage of doing so is the availability of mature eigensolvers for use to solve the underlying QEP. Independently, QEPmin (2.18) is of interest of its own, e.g., it plays a role in solving the total least square problems [20, 37].

3.3.1 Dimensional reduction of QEPmin

The Lanczos process is natural as a method to solve QEP (2.18b) for its leftmost eigenvalue and the corresponding eigenvector. For convenience, we restate QEP (2.18b) here:

$$(PAP-\lambda I)^2 z=\gamma^{-2}b_0b_0^{\mathrm{T}}z,\quad Pz=z. \tag{3.16}$$

Note that we have added the constraint $Pz=z$ since we are only interested in those eigenvectors $z\in\mathcal{N}(C^{\mathrm{T}})$.

Now we discuss how to perform the dimensional reduction of the QEP (3.16) via the projection onto the Krylov subspace generated by the Lanczos process described in Section 3.1. Let Q_k be the orthogonal matrix and T_k be the tridiagonal matrix generated by k steps of the Lanczos process with the matrix $M=PAP$ and the starting vector b_0. We will again have (3.3), i.e.,

$$PAPQ_k=Q_kT_k+\beta_{k+1}q_{k+1}e_k^{\mathrm{T}}\quad\text{and}\quad Q_k^{\mathrm{T}}b_0b_0^{\mathrm{T}}Q_k=\|b_0\|^2e_1e_1^{\mathrm{T}}. \tag{3.17}$$

By a straightforward calculation, we have

$$\begin{aligned}(PAP-\lambda I)^2 Q_k &= (PAP-\lambda I)\big[Q_k(T_k-\lambda I)+\beta_{k+1}q_{k+1}e_k^{\rm T}\big]\\ &=\big[Q_k(T_k-\lambda I)+\beta_{k+1}q_{k+1}e_k^{\rm T}\big](T_k-\lambda I)+(PAP-\lambda I)\beta_{k+1}q_{k+1}e_k^{\rm T}\\ &=Q_k(T_k-\lambda I)^2+\beta_{k+1}q_{k+1}e_k^{\rm T}(T_k-\lambda I)+\beta_{k+1}(PAP-\lambda I)q_{k+1}e_k^{\rm T} \qquad (3.18)\end{aligned}$$

and

$$\begin{aligned}Q_k^{\rm T}(PAP-\lambda I)^2 Q_k &= (T_k-\lambda I)^2+0+\beta_{k+1}Q_k^{\rm T}(PAP-\lambda I)q_{k+1}e_k^{\rm T}\\ &=(T_k-\lambda I)^2+\beta_{k+1}\big[Q_k(T_k-\lambda I)+\beta_{k+1}q_{k+1}e_k^{\rm T}\big]^{\rm T}q_{k+1}e_k^{\rm T}\\ &=(T_k-\lambda I)^2+\beta_{k+1}^2 e_k e_k^{\rm T}. \qquad (3.19)\end{aligned}$$

By (3.17) and (3.19), naturally one would like to take the reduced QEP (3.16) to be

$$\big[(T_k-\lambda I)^2+\beta_{k+1}^2 e_k e_k^{\rm T}\big]w=\gamma^{-2}\|b_0\|^2 e_1(e_1^{\rm T}w). \qquad (3.20)$$

Unfortunately, this reduced QEP may not have any real eigenvalue, not to mention that the leftmost eigenvalue is guaranteed to be real, as demonstrated by Example 3.1 below. To overcome it, we propose to drop the term $\beta_{k+1}^2 e_k e_k^{\rm T}$ in (3.19) and use the following reduced QEP

$$(T_k-\lambda I)^2 w=\gamma^{-2}\|b_0\|^2 e_1(e_1^{\rm T}w). \qquad (3.21)$$

Since it has the same form as the QEP in pQEPmin (2.26b), the leftmost eigenvalue of the reduced QEP (3.21) is guaranteed to be real by Theorem 2.7.

It can be seen that the corresponding reduced QEPmin (2.18) to QEP (3.21) is given by

$$\text{rQEPmin:}\quad \begin{cases}\min \lambda & (3.22\text{a})\\ \text{s.t. } (T_k-\lambda I)^2 w=\gamma^{-2}\|b_0\|^2 e_1(e_1^{\rm T}w), & (3.22\text{b})\\ \qquad \lambda\in\mathbb{R},\ w\neq 0. & (3.22\text{c})\end{cases}$$

We note that the Lanczos process of PAP on b_0 is the same as, upon a linear transformation by $S_1^{\rm T}$, that of H on g_0 in pQEPmin (2.26). Therefore, rQEPmin (3.22) can be viewed as a reduced-form of pQEPmin (2.26).

Example 3.1. Let $A=\text{diag}(1,2,3,4,5)$, $C=[0.65,1,0.68,1.13,-0.23]^{\rm T}$ and $b=[1]$. The eigenvalues of QEP (2.18b) and (2.18c) in QEPmin, computed by MATLAB, are `0.8333`, `1.6493`, `2.0000`, `2.9916`±`0.2369i`, `3.8786`, `4.8236`, `5.1196`. We see the leftmost eigenvalue is real. Apply the Lanczos process with $k=2$ leads to a 2×2 QEP (3.20) whose eigenvalues are computed to be `1.8124`±`0.4172i` and `3.3714`±`0.2547i`, both are genuine complex numbers! In contrast, the eigenvalues of QEP (3.21) are `1.1429`, `2.2661`, `2.8915`, `4.0672`, all of which are real.

3.3.2 Solving rQEPmin

To solve rQEPmin (3.21), we first linearize it into a linear eigenvalue problem (LEP). The reader is referred to [11, Chapter 1] for many different ways to linearize a general polynomial eigenvalue problem. Our rQEPmin (3.21) takes a rather particular form, and we use similar ideas but slightly different linearization. Specifically, we let $y=(T_k-\lambda I)w$ and $s=\begin{bmatrix} y \\ w \end{bmatrix}$. Then QEP (3.22b) can be converted to the following LEP:

$$\begin{bmatrix} T_k & -\gamma^{-2}\|b_0\|^2 e_1 e_1^{\mathrm{T}} \\ -I & T_k \end{bmatrix} s=\lambda s. \tag{3.23}$$

At this point, one can use a standard eigensolver to find the leftmost real eigenvalue $\mu^{(k)}$ of LEP (3.23) and its corresponding eigenvector $s^{(k)}=\begin{bmatrix} y^{(k)} \\ w^{(k)} \end{bmatrix}$. Subsequently, an approximate optimizer of rQEPmin (3.22) is given by $(\mu^{(k)},w^{(k)})$.

3.3.3 Solving QEPmin

The minimizer $(\mu^{(k)},w^{(k)})$ of rQEPmin (3.22) yields an approximate minimizer of QEPmin (2.18) as

$$(\mu^{(k)},z^{(k)})=(\mu^{(k)},Q_k w^{(k)}). \tag{3.24}$$

The accuracy of this pair $(\mu^{(k)},z^{(k)})$ as an approximate minimizer can be measured by the norm of the following the residual vector

$$r_k^{\text{QEPmin}}=\left(PAP-\mu^{(k)}I\right)^2 z^{(k)}-\gamma^{-2}b_0(b_0^{\mathrm{T}}z^{(k)}). \tag{3.25}$$

The following proposition shows that this residual vector can be efficiently obtained during computation.

Proposition 3.2. *Suppose that $(\mu^{(k)},w^{(k)})$ is an exact minimizer of* rQEPmin (3.22) *and $y^{(k)}=(T_k-\mu^{(k)}I)w^{(k)}$. Then*

$$r_k^{\text{QEPmin}}=\beta_{k+1}q_{k+1}e_k^{\mathrm{T}}y^{(k)}+\beta_{k+1}(PAP-\mu^{(k)}I)q_{k+1}(e_k^{\mathrm{T}}w^{(k)}). \tag{3.26}$$

Proof. Keeping (3.18) in mind, we find that

$$\begin{aligned} r_k^{\text{QEPmin}} &= \left(PAP-\mu^{(k)}I\right)^2 Q_k w^{(k)}-\gamma^{-2}b_0 b_0^{\mathrm{T}} Q_k w^{(k)} \\ &\overset{(3.18)}{=} Q_k(T_k-\mu^{(k)}I)^2 w^{(k)}+\beta_{k+1}q_{k+1}e_k^{\mathrm{T}}(T_k-\mu^{(k)}I)w^{(k)} \\ &\qquad +\beta_{k+1}(PAP-\mu^{(k)}I)q_{k+1}e_k^{\mathrm{T}}w^{(k)}-Q_k\frac{\|b_0\|^2}{\gamma^2}e_1(e_1^{\mathrm{T}}w^{(k)}) \\ &\overset{(3.22b)}{=} \beta_{k+1}q_{k+1}e_k^{\mathrm{T}}(T_k-\mu^{(k)}I)w^{(k)}+\beta_{k+1}(PAP-\mu^{(k)}I)q_{k+1}(e_k^{\mathrm{T}}w^{(k)}) \\ &= \beta_{k+1}q_{k+1}e_k^{\mathrm{T}}y^{(k)}+\beta_{k+1}(PAP-\mu^{(k)}I)q_{k+1}(e_k^{\mathrm{T}}w^{(k)}), \end{aligned}$$

as expected. □

We note that if the $(k+1)$st step are carried out in the Lanczos process (3.3), then the term $(PAP-\mu^{(k)}I)q_{k+1}$ in (3.26) can be expressed as a linear combination of q_k, q_{k+1}, and q_{k+2}. We propose to use the following normalized residual norm as a stopping criterion for the Lanczos process:

$$\begin{aligned}\mathrm{NRes}_k^{\mathrm{QEPmin}} &:= \frac{\|r_k^{\mathrm{QEPmin}}\|}{\big[(\|A\|+|\mu^{(k)}|)^2+\gamma^{-2}\|b_0\|^2\big]\|w^{(k)}\|_2} && \text{(3.27a)}\\ &\leqslant \frac{|\beta_{k+1}|\big[|e_k^{\mathrm{T}}y^{(k)}|+(\|A\|+|\mu^{(k)}|)\,|e_k^{\mathrm{T}}w^{(k)}|\big]}{\big[(\|A\|+|\mu^{(k)}|)^2+\gamma^{-2}\|b_0\|^2\big]\|w^{(k)}\|_2} =: \delta_k^{\mathrm{QEPmin}}. && \text{(3.27b)}\end{aligned}$$

The Lanczos algorithm for solving QEPmin (2.18) is summarized in Algorithm 3.

Algorithm 3 Solving QEPmin (2.18)

```
Input: $A\in\mathbb{R}^{n\times n}$, $C\in\mathbb{R}^{n\times m}$, $b_0\in\mathbb{R}^n$, $\gamma>0$, and tolerance $\epsilon$;
Output: $(\mu^{(k)},z^{(k)})$, approximate minimizer of QEPmin (2.18)
 1: $\beta_1\leftarrow\|b_0\|$;
 2: if $\beta_1=0$ then stop;
 3: $q_1\leftarrow b_0/\beta_1$, $q_0\leftarrow 0$;
 4: for $k=1,2,\cdots$ do
 5:     $\hat{q}\leftarrow Aq_k$, $\hat{q}\leftarrow P\hat{q}$, $\hat{q}\leftarrow\hat{q}-\beta_k q_{k-1}$;
 6:     $\alpha_k\leftarrow q_k^{\mathrm{T}}\hat{q}$, $\hat{q}\leftarrow\hat{q}-\alpha_k q_k$, $\beta_{k+1}\leftarrow\|\hat{q}\|$;
 7:     compute the leftmost eigenpair $(\mu^{(k)},s)$ of LEP (3.23);
 8:     $y^{(k)}\leftarrow s_{(1:k)}$, $w^{(k)}\leftarrow s_{(k+1:2k)}$;
 9:     if $\delta_k^{\mathrm{QEPmin}}\leqslant\epsilon$ then stop;
10:     $q_{k+1}\leftarrow\hat{q}/\beta_{k+1}$;
11: end for
12: $Q_k=[q_1,q_2,\cdots,q_k]$;
13: $z^{(k)}=Q_kw^{(k)}$ and $u^{(k)}=-\frac{\gamma^2}{\|b_0\|e_1^{\mathrm{T}}w^{(k)}}Q_ky^{(k)}$;
14: return $(\mu^{(k)},z^{(k)})$ as an approximated minimizer of QEPmin (2.18) and, as a by-
    product, $(\mu^{(k)},u^{(k)})$ as an approximated minimizer of LGopt (2.13).
```

It remains to explain why $(\mu^{(k)},u^{(k)})$ at line 14 of Algorithm 3 is an approximated minimizer of LGopt (2.13). Let $\left(\mu^{(k)},\begin{bmatrix}y^{(k)}\\ w^{(k)}\end{bmatrix}\right)$ be the leftmost eigenpair of LEP (3.23). By Theorem 3.2, $\mu^{(k)}\notin\mathrm{eig}(T_k)$, $(T_k-\mu^{(k)}I)^2w^{(k)}\neq 0$ and $e_1^{\mathrm{T}}w^{(k)}\neq 0$. Through a straightforward application of Theorem 2.5 to rLGopt (3.5) and rQEPmin (3.22), we find that $(\mu^{(k)},x^{(k)})$ is the minimizer of rLGopt (3.5) where

$$x^{(k)}=-\frac{\gamma^2}{\|b_0\|e_1^{\mathrm{T}}w^{(k)}}(T_k-\mu^{(k)}I)w^{(k)}=-\frac{\gamma^2}{\|b_0\|e_1^{\mathrm{T}}w^{(k)}}y^{(k)}. \tag{3.28}$$

Therefore, as a by-product, an approximate minimizer of LGopt (2.13) is given by

$$(\mu^{(k)},u^{(k)})=\left(\mu^{(k)},-\frac{\gamma^2}{\|b_0\|e_1^{\mathrm{T}}w^{(k)}}Q_ky^{(k)}\right). \tag{3.29}$$

3.4 Lanczos algorithm for CRQopt

Having obtained approximate minimizers of LGopt (2.13) and QEPmin (2.18), by Theorem 2.2 we can recover an approximate minimizer of CRQopt (1.1) as

$$v^{(k)}=n_0+u^{(k)}, \tag{3.30}$$

where $u^{(k)}$ is given by (3.11) if via solving LGopt (2.13) or by (3.29) if via solving QEPmin (2.18). The overall algorithm called *the Lanczos Method*, is outlined in Algorithm 4. In line 6, we can solve LGopt (2.13) by Algorithm 2 for rLGopt (3.5) or Algorithm 3 for rQEPmin (3.22). Since the most time-consuming step is the Lanczos iterations, there is no significant difference between two algorithms in overall efficiency. We include them for the sake of completeness in addition to the different advantages of each algorithm discussed in the previous sections.

Algorithm 4 Solving CRQopt (1.1)

```
Input: $A\in\mathbb{R}^{n\times n}$, $C\in\mathbb{R}^{n\times m}$ with full column rank, $b\in\mathbb{R}^m$, tolerance $\epsilon$;
Output: approximate minimizer $v$ of CRQopt (1.1);
1: $n_0\leftarrow(C^{\mathrm{T}})^{\dagger}b$ (by, e.g., the QR decomposition of $C$);
2: if $\|n_0\|>1$ then output no solution;
3: if $\|n_0\|=1$ then $v\leftarrow n_0$ and output $v$;
4: if $\|n_0\|<1$ then
5:     $\gamma\leftarrow\sqrt{1-\|n_0\|^2}$, $q\leftarrow An_0$, $b_0\leftarrow(I-CC^{\dagger})q$;
6:     compute an approximate solution of LGopt (2.13) $(\mu^{(k)},u^{(k)})$ by Algorithm 2 or 3
7:     return $v^{(k)}=n_0+u^{(k)}$, approximate minimizer of CRQopt (1.1);
8: end if
```

3.4.1 Finite step stopping property

As in many Lanczos type methods for numerical linear algebra problems [4,13,30,34], Algorithm 4 enjoys a finite-step-stopping property in the exact arithmetic, i.e., it will deliver an exact solution in at most n steps. It is an excellent theoretic property but of little practical significance for large scale problems. We often expect that the Lanczos process would stop much sooner before the nth step for otherwise the method would be deemed too expensive to be practical.

We will show the property using LGopt (2.13) as an example, which, for convenience, is restated here:

$$\text{LGopt:}\quad \begin{cases} \min \lambda & \text{(2.13a)}\\ \text{s.t. } (PAP-\lambda I)u=-b_0, & \text{(2.13b)}\\ \quad \|u\|=\gamma, & \text{(2.13c)}\\ \quad u\in\mathcal{N}(C^{\mathrm{T}}). & \text{(2.13d)} \end{cases}$$

Let (λ_*,u_*) be the minimizer of LGopt (2.13) and $k_{\max}$ be the smallest k such that $\beta_{k+1}=0$ in the Lancozs process, namely the Lanczos process breaks down at step $k=k_{\max}$. We will prove that $\mu^{(k_{\max})}=\lambda_*$ and $u^{(k_{\max})}=u_*$.

We have already shown in (3.12) that the second and third constraints of LGopt (2.13) are satisfied by $u^{(k_{\max})}$. Besides, since $\beta_{k_{\max}+1}=0$, $r^{\rm LGopt}_{k_{\max}}=0$ by Proposition 3.1, i.e., the first constraint of LGopt (2.13) holds. It remains to show that $\mu^{(k_{\max})}=\lambda_*$.

Lemma 3.1. *$\mu^{(k_{\max})}$ is the smallest root of*

$$\widetilde{\chi}(\lambda):=g^{\mathrm{T}}[(H-\lambda I)^{\dagger}]^2g^{\mathrm{T}}-\gamma^2. \tag{3.31}$$

In addition, if LGopt (2.13) *is in the easy case, then $\mu^{(k_{\max})}=\lambda_*$, where (λ_*,z_*) is the minimizer of* LGopt (2.13).

Proof. Let $\vartheta_1\leqslant\vartheta_2\leqslant\cdots\leqslant\vartheta_{k_{\max}}$ be the eigenvalues of $T_{k_{\max}}$ and let $y_1,y_2,\cdots,y_{k_{\max}}$ be the corresponding orthonormal eigenvectors. Expand $\|b_0\|e_1=\sum_{i=1}^{k_{\max}}\zeta_iy_i$ and define the secular function

$$\widehat{\chi}(\lambda)=\|b_0\|^2e_1^{\mathrm{T}}(T_{k_{\max}}-\lambda I)^{-2}e_1-\gamma^2=\sum_{i=1}^{k_{\max}}\frac{\zeta_i^2}{(\lambda-\vartheta_i)^2}-\gamma^2. \tag{3.32}$$

By Theorem 3.2, $\mu^{(k_{\max})}<\vartheta_1$. Apply Lemma 2.7 with $H=T_{k_{\max}}$ and $g=\|b_0\|e_1$ to conclude that $\mu^{(k_{\max})}$ is a root of the secular function (3.32). Since $\widehat{\chi}(\lambda)$ is strictly increasing in $(-\infty,\mu^{(k_{\max})})$, $\mu^{(k_{\max})}$ is the smallest root of $\widehat{\chi}(\lambda)$.

Expand $Q_{k_{\max}}$ to form an the orthogonal matrix $\widehat{Q}:=[Q_{k_{\max}},\ Q_\perp]\in\mathbb{R}^{n\times n}$ and let $T=\widehat{Q}^{\mathrm{T}}PAP\widehat{Q}$. Since the column space of $Q_{k_{\max}}$ is an invariant subspace of PAP, we have

$$T=\begin{bmatrix} T_{k_{\max}} & \\ & T_\perp \end{bmatrix}.$$

Let $S=[S_1,S_2]$ be defined in (2.19), and let $H=S_1^{\mathrm{T}}PAPS_1$ and $g_0=S_1^{\mathrm{T}}b_0$. For any $\lambda<\vartheta_1$, we have

$$\begin{aligned}\widehat{\chi}(\lambda)&=\|b_0\|e_1^{\mathrm{T}}[(T_{k_{\max}}-\lambda I)^{-1}]^2\|b_0\|e_1-\gamma^2\\&=\|b_0\|e_1^{\mathrm{T}}[(T-\lambda I)^{\dagger}]^2\|b_0\|e_1-\gamma^2\\&=b_0^{\mathrm{T}}\widehat{Q}\widehat{Q}^{\mathrm{T}}[(PAP-\lambda I)^{\dagger}]^2\widehat{Q}\widehat{Q}^{\mathrm{T}}b_0-\gamma^2\\&=b_0^{\mathrm{T}}[(PAP-\lambda I)^{\dagger}]^2b_0-\gamma^2\\&=b_0^{\mathrm{T}}SS^{\mathrm{T}}[(PAP-\lambda I)^{\dagger}]^2SS^{\mathrm{T}}b_0-\gamma^2\\&=[g_0^{\mathrm{T}}\ 0]\begin{bmatrix}[(H-\lambda I)^{\dagger}]^2 & 0\\ 0 & [(-\lambda I)^{\dagger}]^2\end{bmatrix}[g_0^{\mathrm{T}}\ 0]^{\mathrm{T}}-\gamma^2\\&=g_0^{\mathrm{T}}[(H-\lambda I)^{\dagger}]^2g_0-\gamma^2=:\widetilde{\chi}(\lambda).\end{aligned}$$

Therefore, $\widetilde{\chi}(\lambda)=0$ and $\widetilde{\chi}(\lambda)<0$ for $\lambda<\mu^{(k_{\max})}$, implying $\mu^{(k_{\max})}$ is the smallest root of $\widetilde{\chi}(\lambda)$.

On the other hand, by the definition of the easy case, $b_0^{\mathrm{T}}z_*\neq 0$ for all possible minimizers (λ_*,z_*) of QEPmin (2.18). Theorem 2.4 says that $z_*=S_1w_*$ for some $w_*\in\mathbb{R}^{n-m}$ and thus $g^{\mathrm{T}}w_*=b_0^{\mathrm{T}}S_1w_*=b_0^{\mathrm{T}}z_*\neq 0$. By Theorem 2.6, $\lambda_*<\lambda_{\min}(H)$. Therefore, it is related to case (1) or subcase (i) in case (2) of the proof in Lemma 2.4, for which λ_* is the smallest root of $\widetilde{\chi}(\lambda)$, and thus $\lambda_*=\mu^{(k_{\max})}$. □

Theorem 2.13 guarantees that the minimizer of CRQopt (1.1) is unique if QEPmin (2.18) is in the easy case. We also have established a finite step stopping property for Algorithm 4 as detailed in the following theorem, since $k_{\max}\leqslant n$.

Corollary 3.1. *Suppose* QEPmin (2.18) *is in the easy case, and let* $(\mu^{(k)},w^{(k)})$ *be the minimizer of* rQEPmin (3.22). *Define* $u^{(k)}$ *as in* (3.11) *and* $k_{\max}$ *is the smallest* k *such that* $\beta_{k+1}=0$. *Then* $\big(\mu^{(k_{\max})},u^{(k_{\max})}\big)$ *solves* LGopt (2.13), *and* $v^{(k_{\max})}=u^{(k_{\max})}+n_0$ *is the unique minimizer of* CRQopt (1.1).

3.4.2 Hard case

The hard case is characterized by Theorem 2.12 and we translate $g_0\perp\mathcal{U}$ into $b_0\perp\mathcal{V}$, where $\mathcal{V}$ is the eigenspace of PAP associated with its eigenvalue $\lambda_{\min}(H)$. For this reason, $\mathcal{K}_k(PAP,b_0)$ will contain no eigen-information of PAP associated with $\lambda_{\min}(H)$. Nonetheless, rLGopt (3.5) and rQEPmin (3.22) can be still formed and solved to yield approximations to the original CRQopt (1.1) with suitable stoping criteria satisfied. But the approximations will be utterly wrong if it is indeed in the hard case. Hence in practice it is important to detect when the hard case occurs.

Denote by (λ_*,z_*) the minimizer of LGopt (2.13). In the easy case, the smallest root of $\widetilde{\chi}(\lambda)$ is λ_* and $\lambda_*<\lambda_{\min}(H)$, while in the hard case, $\lambda_*=\lambda_{\min}(H)$ and the smallest root of $\widetilde{\chi}(\lambda)$ defined in (3.31) is greater than or equal to $\lambda_{\min}(H)$. Since $\mu^{(k)}$ converges to $\mu^{(k_{\max})}$, eventually whether $\mu^{(k)}<\lambda_{\min}(H)$ provide a reasonably good test to see if it is the easy case. Therefore, we propose to detect the hard case as follows:

1. Solve rLGopt (3.5) or rQEPmin (3.22);

2. Run the Lanczos process with $M=PAP$ with $r_0=Pc$, where $c\in\mathbb{R}^n$ is random to compute $\lambda_{\min}(H)$ of PAP and its associated eigenvector $\tilde{z}$;

3. Check if the optimal value of rLGopt (3.5) or rQEPmin (3.22) is greater than or equal to $\lambda_{\min}(H)$ within a prescribed accuracy;

4. If it is, then QEPmin (2.18) is in the hard case. Compute an approximation $\tilde{x}$ of $x_*=-(PAP-\lambda_*I)^{\dagger}b_0$
$$\tilde{y}=\operatorname*{argmin}_{y\in\mathbb{R}^k}\left\|\begin{bmatrix} T_k \\ \beta_{k+1}e_k^{\mathrm{T}}\end{bmatrix}y+\|b_0\|e_1\right\|, \quad \tilde{x}=Q_k\tilde{y},$$
and then an approximate minimizer of LGopt (2.13) is given by $\tilde{x}+\sqrt{\gamma^2-\|\tilde{x}\|^2}(\tilde{z}/\|\tilde{z}\|)$.

A remark is in order for item 2 above. Because of the randomness in c, with probability 1, $r_0=Pc$ will have a significant component in $S_1\mathcal{U}$, where $\mathcal{U}$ is as defined in Theorem 2.11.

3.5 Convergence analysis

In this section, we present a convergence analysis of the Lanczos algorithm (Algorithm 4) for solving CRQopt (1.1) in the easy case. Let $h(v)=v^{\mathrm{T}}Av$ be the objective function of CRQopt (1.1), v_* be the unique solution of CRQopt (1.1) and (λ_*,u_*) be the solution of LGopt (2.13). Our main results are upper bounds on the errors $h(v^{(k)})-h(v_*)$, $\|v^{(k)}-v_*\|$ and $|\mu^{(k)}-\lambda_*|$, where $v^{(k)}$ defined in (3.30) is the kth approximation by Algorithm 4 and $(\mu^{(k)},x^{(k)})$ is the solution of rLGopt (3.5).

Our analysis is analogous to the one in [43]. We start by establishing an optimality property of $v^{(k)}$, as an approximation of v_*, that minimizes $h(v)$ over $n_0+\mathcal{K}_k(PAP,b_0)$.

Theorem 3.3. *Let $v^{(k)}$ be defined in* (3.30)*. Then it holds that*
$$h(v^{(k)})=\min_{v\in n_0+\mathcal{K}_k(PAP,b_0),\|v\|=1} h(v). \tag{3.33}$$

Proof. Recall that $(\mu^{(k)},x^{(k)})$ solves rLGopt (3.5). Consider the optimization problem
$$\begin{cases}\min\ \ell(x):=x^{\mathrm{T}}T_kx+2\|b_0\|e_1^{\mathrm{T}}x, & \text{(3.34a)}\\ \ \text{s.t. } \|x\|=\gamma. & \text{(3.34b)}\end{cases}$$

By the theory of Lagrangian multipliers, we find the Lagrangian equations for (3.34) are
$$(T_k-\lambda I)x=-\|b_0\|e_1, \quad \|x\|=\gamma. \tag{3.35}$$

Following the same argument as we did to prove Lemma 2.1, we can reach the same conclusion that $\ell(x)$ is strictly increasing with respect to λ in the solution pair (λ,x) of (3.35). Therefore, in order to minimize $\ell(x)$, we need to find the smallest Lagrangian multiplier satisfying (3.35). Hence, solving (3.34) is equivalent to solving rLGopt (3.5) for which $(\mu^{(k)},x^{(k)})$ is a minimizer and thus $x^{(k)}$ solves (3.34), where $x^{(k)}$ is defined in (3.28).

By definition, $u^{(k)}=Q_kx^{(k)}$ and $v^{(k)}=u^{(k)}+n_0$. For any $v\in n_0+\mathcal{K}_k(PAP,b_0)$ with $\|v\|=1$, let

$$u=v-n_0\in\mathcal{K}_k(PAP,b_0)\subset\mathcal{N}(C^{\mathrm{T}}). \tag{3.36}$$

Hence $Pu=u$, $\|u\|=\gamma$, and $u=Q_k\widetilde{u}$ for some $\widetilde{u}\in\mathbb{R}^k$. We have $v=u+n_0=Pu+n_0$ and

$$\begin{aligned}
h(v)&=(Pu+n_0)^{\mathrm{T}}A(Pu+n_0)\\
&=u^{\mathrm{T}}PAPu+2b_0^{\mathrm{T}}u+n_0^{\mathrm{T}}An_0\\
&=\widetilde{u}^{\mathrm{T}}Q_k^{\mathrm{T}}PAPQ_k\widetilde{u}+2b_0^{\mathrm{T}}Q_k\widetilde{u}+n_0^{\mathrm{T}}An_0\\
&=\widetilde{u}^{\mathrm{T}}T_k\widetilde{u}+2\|b_0\|e_1^{\mathrm{T}}\widetilde{u}+n_0^{\mathrm{T}}An_0\\
&\geqslant[x^{(k)}]^{\mathrm{T}}T_kx^{(k)}+2\|b_0\|e_1^{\mathrm{T}}x^{(k)}+n_0^{\mathrm{T}}An_0 \qquad (\text{since } x^{(k)} \text{ solves (3.34)})\\
&=[x^{(k)}]^{\mathrm{T}}Q_k^{\mathrm{T}}PAPQ_kx^{(k)}+2b_0^{\mathrm{T}}Q_kx^{(k)}+n_0^{\mathrm{T}}An_0\\
&=[u^{(k)}]^{\mathrm{T}}PAPu^{(k)}+2b_0^{\mathrm{T}}u^{(k)}+n_0^{\mathrm{T}}An_0\\
&=(u^{(k)}+n_0)^{\mathrm{T}}A(u^{(k)}+n_0)\\
&=h(v^{(k)}).
\end{aligned}$$

Since $v\in n_0+\mathcal{K}_k(PAP,b_0)$ with $\|v\|=1$ but otherwise is arbitrary, (3.33) holds. □

Recall that H and g_0 are defined in (2.21) and S_1, S_2 in (2.19). Let $\theta_{\min}$ and $\theta_{\max}$ be the smallest and the largest eigenvalue of H, respectively, v_* be the minimizer of CRQopt (1.1), and λ_* be the optimal objective value of LGopt (2.13). Then

$$(\lambda_*,u_*) \quad \text{with} \quad u_*=Pv_*=v_*-n_0$$

is a minimizer of LGopt (2.13). Set

$$\kappa\equiv\kappa(H-\lambda_*I):=\frac{\theta_{\max}-\lambda_*}{\theta_{\min}-\lambda_*}.$$

To estimate $h(v^{(k)})-h(v_*)$, $\|v^{(k)}-v_*\|$ and $|\mu^{(k)}-\lambda_*|$, we first establish a lemma that provides a way to bound $h(v^{(k)})-h(v_*)$, $\|v^{(k)}-v_*\|$ and $|\mu^{(k)}-\lambda_*|$ in terms of any nonzero $v\in n_0+\mathcal{K}_k(PAP,b_0)$.

Lemma 3.2. *For any nonzero $v\in n_0+\mathcal{K}_k(PAP,b_0)$, we have*

$$0\leqslant h(v^{(k)})-h(v_*)\leqslant 4\|H-\lambda_*I\|_2\cdot\|v-v_*\|_2^2, \tag{3.37a}$$

$$\|v^{(k)}-v_*\|\leqslant 2\sqrt{\kappa}\|v-v_*\|_2, \tag{3.37b}$$

$$|\mu^{(k)}-\lambda_*|\leqslant\frac{1}{\gamma^2}\big[4\|H-\lambda_*I\|_2\cdot\|v-v_*\|_2^2+2\sqrt{\kappa}\|b_0\|_2\cdot\|v-v_*\|_2\big]. \tag{3.37c}$$

Proof. For $v\in n_0+\mathcal{K}_k(PAP,b_0)$, let

$$u=v-n_0\in\mathcal{K}_k(PAP,b_0),\quad \widetilde{u}=\gamma u/\|u\|,\quad \widetilde{v}=n_0+\widetilde{u}\in n_0+\mathcal{K}_k(PAP,b_0). \tag{3.38}$$

First, we have $|\|u\|-\gamma|=|\|u\|-\|u_*\||\leqslant\|u-u_*\|=\|v-v_*\|$, which leads to

$$\left|1-\frac{\gamma}{\|u\|}\right|\leqslant\frac{\|v-v_*\|}{\|u\|}. \tag{3.39}$$

Let $r=\widetilde{v}-v_*$, we have

$$\begin{aligned}\|r\|&=\|v_*-\widetilde{v}\|\leqslant\|v_*-v\|+\|v-\widetilde{v}\|\leqslant\|v_*-v\|+\|u-\widetilde{u}\|\\&=\|v_*-v\|+\left\|u-\frac{\gamma u}{\|u\|}\right\|=\|v_*-v\|+\|u\|\times\left|1-\frac{\gamma}{\|u\|}\right|\leqslant2\|v_*-v\|,\end{aligned} \tag{3.40}$$

where we have used (3.39) to infer the last inequality.

The first inequality in (3.37a) holds because

$$h(v^{(k)})=\min_{v\in n_0+\mathcal{K}_k(PAP,b_0),\|v\|=1}h(v)\geqslant\min_{v\in n_0+\mathcal{N}(C^{\rm T}),\|v\|=1}h(v)=h(v_*).$$

Let $f(u)=u^{\rm T}Au+2u^{\rm T}b_0$, it can be verified that $h(v)=h(u+n_0)=f(u)+n_0^{\rm T}An_0$. Therefore,

$$\widetilde{u}-u_*=\widetilde{v}-v_*=r,\quad h(\widetilde{v})-h(v_*)=f(\widetilde{u})-f(u_*). \tag{3.41}$$

Set $s=S_1^{\rm T}r$. It follows from $r\in\mathcal{N}(C^{\rm T})$ that $r=S_1s$ and $\|s\|=\|r\|$. Noting that $\widetilde{v}$ satisfies the constraint of CRQopt (1.1) and that $\widetilde{u}=u_*+r$, we have

$$\begin{aligned}0\leqslant h(v^{(k)})-h(v_*)&\leqslant h(\widetilde{v})-h(v_*)\overset{(3.41)}{=}f(\widetilde{u})-f(u_*)=f(u_*+r)-f(u_*)\\&=r^{\rm T}PAPr+2r^{\rm T}(PAPu_*+b_0)\\&=r^{\rm T}PAPr+2\lambda_*r^{\rm T}u_* &(3.42)\\&=r^{\rm T}(PAP-\lambda_*I)r &(3.43)\\&=s^{\rm T}S_1^{\rm T}(PAP-\lambda_*I)S_1s\\&=s^{\rm T}(H-\lambda_*I)s\\&\leqslant\|H-\lambda_*I\|\|s\|^2=\|H-\lambda_*I\|\|r\|^2\\&\overset{(3.40)}{\leqslant}4\|H-\lambda_*I\|\|v_*-v\|^2, &(3.44)\end{aligned}$$

yielding the second inequality in (3.37a), where we have used $(PAP-\lambda_*I)u_*=-b_0$ to get (3.42) and

$$\|r\|^2+2r^{\rm T}u_*=\|u_*+r\|^2-\|u_*\|^2=\|\widetilde{u}\|^2-\|u_*\|^2=0$$

to obtain $2r^{\rm T}u_*=-r^{\rm T}r$ and then (3.43).

Next we prove (3.37b). Define

$$\widetilde{f}(u) := f(u) - \lambda_* u^{\mathrm{T}} u = u^{\mathrm{T}}(PAP - \lambda_* I)u + 2u^{\mathrm{T}} b_0.$$

Noticing $(PAP-\lambda_* I)u_* + b_0 = 0$ by (2.13b), let $u^{(k)} = v^{(k)} - n_0$, we have

$$\widetilde{f}(u^{(k)}) = \widetilde{f}(u_*) + (u^{(k)} - u_*)^{\mathrm{T}}(PAP - \lambda_* I)(u^{(k)} - u_*).$$

Therefore

$$\widetilde{f}(u^{(k)}) - \widetilde{f}(u_*) \geqslant (\theta_{\min} - \lambda_*)\|u^{(k)} - u_*\|^2 = (\theta_{\min} - \lambda_*)\|v^{(k)} - v_*\|^2.$$

On the other hand,

$$\begin{aligned}\widetilde{f}(u^{(k)}) - \widetilde{f}(u_*) &= [f(u^{(k)}) - \lambda_*\|u^{(k)}\|^2] - [f(u_*) - \lambda_*\|u_*\|^2] \\ &= f(u^{(k)}) - f(u_*) = h(v^{(k)}) - h(v_*),\end{aligned}$$

yielding

$$(\theta_{\min} - \lambda_*)\|v^{(k)} - v_*\|^2 \leqslant h(v^{(k)}) - h(v_*) \leqslant 4\|H - \lambda_* I\|\|v - v_*\|^2, \tag{3.45}$$

which leads to (3.37b).

To prove (3.37c), we pre-multiply $(PAP - \lambda_* I)u_* = -b_0$ by u_*^{T} and use $u_*^{\mathrm{T}} u_* = \gamma^2$ to get

$$\gamma^2 \lambda_* = u_*^{\mathrm{T}} PAP u_* + u_*^{\mathrm{T}} b_0 = v_*^{\mathrm{T}} PAP v_* + v_*^{\mathrm{T}} b_0, \tag{3.46}$$

since $Pv_* = u_*$ and $Pb_0 = b_0$. By (2.4a), we have $h(v_*) = v_*^{\mathrm{T}} PAP v_* + 2v_*^{\mathrm{T}} b_0 + n_0^{\mathrm{T}} A n_0$ and thus

$$\gamma^2 \lambda_* = h(v_*) - v_*^{\mathrm{T}} b_0 - n_0^{\mathrm{T}} A n_0.$$

On the other hand, it follows from rLGopt (3.5) that $[x^{(k)}]^{\mathrm{T}} T_k x^{(k)} + \|b_0\|_2 [x^{(k)}]^{\mathrm{T}} e_1 = \gamma^2 \mu^{(k)}$. Plug in

$$T_k = Q_k^{\mathrm{T}} PAP Q_k, \quad u^{(k)} = Q_k x^{(k)}, \quad Q_k^{\mathrm{T}} b_0 = \|b_0\|_2 e_1, \quad v^{(k)} = u^{(k)} + n_0$$

to get

$$\gamma^2 \mu^{(k)} = h(u^{(k)}) + [u^{(k)}]^{\mathrm{T}} b_0 = h(v^{(k)}) - [v^{(k)}]^{\mathrm{T}} b_0 - n_0^{\mathrm{T}} A n_0. \tag{3.47}$$

It follows from (3.46) and (3.47) that

$$\begin{aligned}\left|\mu^{(k)} - \lambda_*\right| &= \frac{1}{\gamma^2}\left|h(v^{(k)}) - h(v_*) - b_0^{\mathrm{T}}(v^{(k)} - v_*)\right| \\ &\leqslant \frac{1}{\gamma^2}\Big[|h(v^{(k)}) - h(v_*)| + \|b_0\|_2 \|v^{(k)} - v_*\|_2\Big],\end{aligned} \tag{3.48}$$

which combined with (3.37a) and (3.37b) yield (3.37c). □

The inequalities in (3.37) hold for any $v\in n_0+\mathcal{K}_k(PAP,b_0)$ which, in general can be expressed as

$$v=n_0+\phi_{k-1}(PAP)b_0,$$

where $\phi_{k-1}(\cdot)$ is a polynomial of degree $k-1$. By judicially picking certain ϕ_{k-1}, meaningful upper bounds on $h(v^{(k)})-h(v_*)$, $\|v^{(k)}-v_*\|$ and $|\mu^{(k)}-\lambda_*|$ are readily obtained. These upper bounds expose the convergence behavior of $v^{(k)}$. The next theorem contains our main results of the section.

Theorem 3.4. *Suppose* CRQopt (1.1) *is in the easy case, and let v_* be its minimizer. Let (λ_*,u_*) be the minimizer of the corresponding* LGopt (2.13)*, and, for its corresponding* pLGopt (2.23)*, let $\theta_{\min}$ and $\theta_{\max}$ be the smallest and largest eigenvalue of H, respectively, and set*

$$\kappa=\kappa(H-\lambda_*I):=\frac{\theta_{\max}-\lambda_*}{\theta_{\min}-\lambda_*}.$$

Then the following statements hold:

(a) *The sequence $\{h(v^{(k)})\}$ is nonincreasing;*

(b) *For $k\leqslant k_{\max}$, the smallest k such that $\beta_{k+1}=0$,*

$$0\leqslant h(v^{(k)})-h(v_*)\leqslant 16\gamma^2\|H-\lambda_*I\|_2\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}, \tag{3.49a}$$

$$\|v^{(k)}-v_*\|_2\leqslant 4\gamma\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}, \tag{3.49b}$$

$$|\mu^{(k)}-\lambda_*|\leqslant 16\|H-\lambda_*I\|_2\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}+\frac{4}{\gamma}\|b_0\|_2\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}, \tag{3.49c}$$

where

$$\Gamma_\kappa:=\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}. \tag{3.50}$$

Proof. Item (a) holds because for any $0\leqslant k\leqslant k_{\max}$,

$$h(v^{(k)})=\min_{v\in n_0+\mathcal{K}_k(PAP,b_0),\|v\|=1}h(v)\geqslant\min_{v\in n_0+\mathcal{K}_{k+1}(PAP,b_0),\|v\|=1}h(v)=h(v^{(k+1)}).$$

Before we prove item (b), we note that $(\lambda_*,S_1^{\mathrm{T}}v_*)$ solves pLGopt (2.23). In particular, since pLGopt (2.23) is in the easy case,

$$S_1^{\mathrm{T}}v_*=-(H-\lambda_*I)^{-1}g_0. \tag{3.51}$$

Consider now $v\in n_0+\mathcal{K}_k(PAP,b_0)$. Then $S_1^{\mathrm{T}}v\in\mathcal{K}_k(H,g_0)=\mathcal{K}_k(H-\lambda_*I,g_0)$. Therefore by (3.51)

$$\begin{aligned}S_1^{\mathrm{T}}v-S_1^{\mathrm{T}}v_*&=\phi_{k-1}(H-\lambda_*I)g+(H-\lambda_*I)^{-1}g_0\\&=[\phi_{k-1}(H-\lambda_*I)(H-\lambda_*I)+I](H-\lambda_*I)^{-1}g_0\\&=-\psi_k(H-\lambda_*I)S_1^{\mathrm{T}}v_*,\end{aligned} \tag{3.52}$$

where ϕ_{k-1} is a polynomial of degree $k-1$, and $\psi_k(t)=1+t\phi_{k-1}(t)$, a polynomial of degree k, that satisfies $\psi_k(0)=1$. Note that $\psi_k(0)=1$ but otherwise ψ_k is an arbitrary polynomial of degree k, offering the freedom that we will take advantage of in a moment.

Given that v_* solves CRQopt (1.1), we have

$$\gamma=\|Pv_*\|=\|S_1S_1^{\mathrm{T}}v_*\|=\|S_1^{\mathrm{T}}v_*\|.$$

Thus

$$\begin{aligned}\min_{v\in n_0+\mathcal{K}_k(PAP,b_0)}\|v-v_*\| &= \min_{v\in n_0+\mathcal{K}_k(PAP,b_0)}\|S_1^{\mathrm{T}}v-S_1^{\mathrm{T}}v_*\| \qquad \text{(use (3.52))}\\ &\leqslant \gamma\min_{\psi_k(0)=1}\|\psi_k(H-\lambda_*I)\|\\ &\leqslant \gamma\min_{\psi_k(0)=1}\max_{1\leqslant i\leqslant n-m}|\psi_k(\theta_i-\lambda_*)| &(3.53)\\ &\leqslant \gamma\min_{\psi_k(0)=1}\max_{t\in[\theta_{\min}-\lambda_*,\theta_{\max}-\lambda_*]}|\psi_k(t)|. &(3.54)\end{aligned}$$

The inequality (3.54) holds for any polynomial ψ_k of degree k such that $\psi_k(0)=1$. For the purpose of establishing upper bounds, we will pick one that is defined through the kth Chebyshev polynomial of the first kind:

$$\begin{aligned}\mathscr{T}_k(t) &= \cos(k\arccos t) &&\text{for } |t|\leqslant 1, &(3.55a)\\ &=\frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^k+\left(t+\sqrt{t^2-1}\right)^{-k}\right] &&\text{for } |t|\geqslant 1. &(3.55b)\end{aligned}$$

Specifically, we take

$$\psi_k(t)=\mathscr{T}_k\left(\frac{2t-(\alpha+\beta)}{\beta-\alpha}\right)\Big/\mathscr{T}_k\left(\frac{-(\alpha+\beta)}{\beta-\alpha}\right), \tag{3.56}$$

where $\alpha=\theta_{\min}-\lambda_*$ and $\beta=\theta_{\max}-\lambda_*$. Evidently, $\psi_k(0)=1$, and for $t\in[\theta_{\min}-\lambda_*,\theta_{\max}-\lambda_*]=[\alpha,\beta]$, we have

$$|2t-(\alpha+\beta)|=||t+\lambda_*-\theta_{\min}|-|t+\lambda_*-\theta_{\max}||\leqslant|\theta_{\max}-\theta_{\min}|=\beta-\alpha.$$

Therefore, $[2t-(\alpha+\beta)]/(\beta-\alpha)\in[-1,1]$, and thus for $t\in[\alpha,\beta]$ [23]

$$|\psi_k(t)|\leqslant\left|\mathscr{T}_k\left(\frac{-(\alpha+\beta)}{\beta-\alpha}\right)\right|^{-1}=\left|\mathscr{T}_k\left(\frac{\kappa+1}{\kappa-1}\right)\right|^{-1}=2\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}. \tag{3.57}$$

Minimize the right-most quantities in (3.37) over $v\in n_0+\mathcal{K}_k(PAP,b_0)$, utilize (3.54) and (3.57) to get the inequalities in (3.49). □

We end this section with remarks regarding the results in Theorem 3.4.

Remark 3.2. The rate of convergence for the Lanczos algorithm depends on κ. Recall that $\kappa=\frac{\theta_{\max}-\lambda_*}{\theta_{\min}-\lambda_*}$. When λ_* is far away from $\theta_{\min}$, we may regard that CRQopt (1.1) is far from hard case. In this case, κ moves towards 1, and we expect faster convergence of our Lanczos algorithm. However, when CRQopt (1.1) is near hard case, i.e., $\theta_{\min}\approx\lambda_*$, κ is large, and Theorem 3.4 suggests slow convergence. These conclusions derived from Theorem 3.4 are consistent with the numerical observations in [17] that "a Lanczos type process seems to be very effective when the problem is far from the hard case". We provide an example in Example 3.3 later to illustrate the relationship between the rate of convergence and κ.

Remark 3.3. The bounds in (3.49a) and (3.49b) are generally sharp. However, there are some cases where the bounds suggested in (3.49a) and (3.49b) are pessimistic. This occurs for near-hard-case situations where $\lambda_*\approx\theta_{\min}$. Although the Lanczos method could still enjoy fast convergence, the bounds in (3.49a) and (3.49b) do not suggest so. One of such situations is when

$$\kappa_+:=\frac{\theta_{\max}-\lambda_*}{\theta_2-\lambda_*}$$

is small, even though $\theta_{\min}\approx\lambda_*$ and thus κ is huge, where θ_2 is the second smallest eigenvalue of H. This suggests that the bounds by (3.49a) and (3.49b) have room for improvement. In fact, instead of (3.56), we may choose

$$\psi_k(t)=\frac{t-\alpha}{-\alpha}\cdot\mathscr{T}_{k-1}\left(\frac{2t-(\alpha_++\beta)}{\beta-\alpha_+}\right)\Big/\mathscr{T}_{k-1}\left(\frac{-(\alpha_++\beta)}{\beta-\alpha_+}\right), \tag{3.58}$$

where α and β are as before, and $\alpha_+=\theta_2-\lambda_*$. Evidently, again $\psi_k(0)=1$, but now $\psi_k(\theta_1-\lambda_*)=0$. We have

$$\begin{aligned}\max_{1\leqslant i\leqslant n-m}|\psi_k(\theta_i-\lambda_*)|=\max_{2\leqslant i\leqslant n-m}|\psi_k(\theta_i-\lambda_*)|&\leqslant\max_{t\in[\alpha_+,\beta]}|\psi_k(t)|\\ &\leqslant\max_{t\in[\alpha_+,\beta]}\left|\frac{t-\alpha}{-\alpha}\right|\cdot 2\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1}\\ &=2\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1}.\end{aligned} \tag{3.59}$$

By combining with (3.53), it leads to the following bounds

$$h(v^{(k)})-h(v_*)\leqslant16\gamma^2\|H-\lambda_*I\|_2\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)^2\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-2}, \tag{3.60a}$$

$$\|v^{(k)}-v_*\|_2\leqslant4\gamma\sqrt{\kappa}\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1}, \tag{3.60b}$$

$$\begin{aligned}|\mu^{(k)}-\lambda_*|\leqslant&16\|H-\lambda_*I\|_2\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)^2\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-2}\\ &+\frac{4}{\gamma}\|b_0\|_2\sqrt{\kappa}\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1}.\end{aligned} \tag{3.60c}$$

These bounds can be much sharper than the ones in (3.49) if $\theta_{\min} \approx \lambda_*$ and there is a reasonably gap between $\theta_{\min}$ and θ_2, see Example 3.4.

Remark 3.4. In our numerical experiments, we observed that the bound (3.49c) often decays much slower than $|\mu^{(k)} - \lambda_*|$. Recall that in obtaining (3.49c), in (3.48), we used the inequality

$$\left| b_0^{\mathrm{T}}(v^{(k)} - v_*) \right| \leqslant \|b_0\| \left\| v^{(k)} - v_* \right\|. \tag{3.61}$$

It turns out that $\|b_0\| \|v^{(k)} - v_*\|$ decays much slower than $|b_0^{\mathrm{T}}(v^{(k)} - v_*)|$, as evidenced by numerical tests. While at this point we do not know how to estimate $|b_0^{\mathrm{T}}(v^{(k)} - v_*)|$ much more accurately than the inequality (3.61), we offer a plausible explanation as follows. Let $u^{(k)} = v^{(k)} - n_0$ and $u_* = v_* - n_0$. Since $u_*^{\mathrm{T}} u_* = [u^{(k)}]^{\mathrm{T}} u^{(k)} = \gamma^2$, we have

$$\begin{aligned}\left| u_*^{\mathrm{T}}(v^{(k)} - v_*) \right| &= \left| u_*^{\mathrm{T}} u^{(k)} - u_*^{\mathrm{T}} u_* \right| = \frac{1}{2} \left| 2u_*^{\mathrm{T}} u^{(k)} - u_*^{\mathrm{T}} u_* - [u^{(k)}]^{\mathrm{T}} u^{(k)} \right| \\ &= \frac{1}{2} \left\| u^{(k)} - u_* \right\|_2^2 = \frac{1}{2} \left\| v^{(k)} - v_* \right\|_2^2. \end{aligned} \tag{3.62}$$

By (3.49b), $\|v^{(k)} - v_*\|_2^2$ is of order $\mathcal{O}\big([\Gamma_\kappa^k + \Gamma_\kappa^{-k}]^{-2}\big)$, and thus $|u_*^{\mathrm{T}}(v^{(k)} - v_*)|$ is also of order $\mathcal{O}\big([\Gamma_\kappa^k + \Gamma_\kappa^{-k}]^{-2}\big)$ as (3.62) suggests. Let $\theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_{n-m}$ be the eigenvalues of PAP restricted to the subspace $\mathcal{R}(P)$, $y_1, y_2, \cdots, y_{n-m}$ be the corresponding orthonormal eigenvectors in $\mathcal{R}(P)$, $u_* = \sum_{i=1}^{n-m} \xi_i y_i$, and $v^{(k)} - v_* = u^{(k)} - u_* = \sum_{i=1}^{n-m} \epsilon_i y_i$. Then we have

$$\left| u_*^{\mathrm{T}}(v^{(k)} - v_*) \right| = \left| \sum_{i=1}^{n-m} \xi_i \epsilon_i \right|.$$

On the other hand, $b_0 = -(PAP - \lambda_* I) u_* = -\sum_{i=1}^{n-m} (\theta_i - \lambda_*) \xi_i y_i$ and thus

$$\left| b_0^{\mathrm{T}}(v^{(k)} - v_*) \right| = \left| \sum_{i=1}^{n} (\theta_i - \lambda_*) \xi_i \epsilon_i \right|.$$

Note that the sequence $\{\theta_i - \lambda_*\}$ is positive and increasing for the easy case and the sequence $\{\xi_i y_i\}$ oscillates. Therefore, when $\kappa(PAP - \lambda_* I) = \frac{\theta_{n-m} - \lambda_*}{\theta_1 - \lambda_*}$ is modest, i.e., the difference between $\theta_i - \lambda_*$ for different i is modest, we expect that the difference between $|b_0^{\mathrm{T}}(v^{(k)} - v_*)| = |\sum_{i=1}^{n-m} (\theta_i - \lambda_*) \xi_i \epsilon_i|$ and $|u_*^{\mathrm{T}}(v^{(k)} - v_*)| = |\sum_{i=1}^{n-m} \xi_i \epsilon_i|$ is small. Therefore, the convergence rate of $|b_0^{\mathrm{T}}(v^{(k)} - v_*)|$ can be similar to the convergence rate of $|u_*^{\mathrm{T}}(v^{(k)} - v_*)|$, which is $\mathcal{O}\big([\Gamma_\kappa^k + \Gamma_\kappa^{-k}]^{-2}\big)$. This explains why the bound (3.49c) decays much slower than $|\mu^{(k)} - \lambda_*|$.

3.6 Numerical examples

In this section, we demonstrate the sharpness of the convergence error bounds in Theorem 3.4 for the Lanczos algorithm (Algorithm 4) for solving CRQopt (1.1). For that purpose, we first provide examples that are considered to be hard for the Lanczos algorithm.

The basic idea is similar to the one in [24]. Also shown are the history of the normalized residual $\mathrm{NRes}_k^{\mathrm{QEPmin}}$ and its upper bound $\delta_k^{\mathrm{QEPmin}}$ in (3.27b).

3.6.1 Construction of difficult CRQopt problems

The convergence analysis of the Lanczos algorithm (Algorithm 4) for solving CRQopt (1.1) presented in Theorem 3.4 indicates that the convergence behavior is determined by the spectral distribution of the matrix H defined in pLGopt (2.23) and the optimal value λ_* of LGopt (2.13). It is not a secret that approximations by the Lanczos procedure converge most slowly when the eigenvalues of H distribute like the zeros or the extreme nodes of Chebyshev polynomials of the first kind [22–24,43]. Therefore, we construct matrices A, C and vector b of CRQopt through constructing matrices H and g_0 of pLGopt (2.23).

In what follows, we describe one set of test matrix-vector pairs (H,g_0) using the extreme nodes of Chebyshev polynomials of the first kind. Recall that the ℓth Chebyshev polynomials of the first kind $\mathscr{T}_\ell(t)$ has $\ell+1$ extreme points in $[-1,1]$, defined by

$$\tau_{j\ell}=\cos\vartheta_{j\ell},\quad \text{with}\quad \vartheta_{jl}=\frac{j}{\ell}\pi \quad \text{for}\quad j=0,1,\cdots,\ell. \tag{3.63}$$

At these extreme points, $|\mathscr{T}_\ell(\tau_{j\ell})|=1$. Given scalars α and β such that $\alpha<\beta$, set

$$\omega=\frac{\beta-\alpha}{2},\quad \tau=-\frac{\alpha+\beta}{\beta-\alpha}. \tag{3.64}$$

The so-called *the ℓth translated Chebyshev extreme nodes* on $[\alpha,\beta]$ are given by [22,23]

$$\tau_{j\ell}^{\mathrm{trans}}=\omega(\tau_{j\ell}-\tau)\quad \text{for}\quad j=0,1,\cdots,\ell. \tag{3.65}$$

It can be verified that $\tau_{0\ell}^{\mathrm{trans}}=\beta$ and $\tau_{\ell\ell}^{\mathrm{trans}}=\alpha$.

Given integers n and m with $m<n$, and the interval $[\alpha,\beta]$, we take

$$H=\mathrm{diag}\left(\tau_{0n-m-1}^{\mathrm{trans}},\tau_{1n-m-1}^{\mathrm{trans}},\cdots,\tau_{n-m-1n-m-1}^{\mathrm{trans}}\right). \tag{3.66}$$

Now we construct $g_0=[g_1,g_2,\cdots,g_{n-m}]^{\mathrm{T}}\in\mathbb{R}^{n-m}$. Recall that the eigenvector of H corresponding to the smallest eigenvalue is e_{n-m}. In order to make pLGopt (2.23) in the easy case, we need to make g_0 not perpendicular to that eigenvector e_{n-m}, i.e., $g_{n-m}\neq 0$. As a simple choice, we take

$$g_0=[1,1,\cdots,1]^{\mathrm{T}}\in\mathbb{R}^{n-m}. \tag{3.67}$$

With H and g_0 set, we construct matrices A, C and vector b in the following way:

1. Pick $0<\zeta<1$, and $a\in\mathbb{R}^m$ with $\|a\|=1/\zeta$;

2. Pick a random $C\in\mathbb{R}^{n\times m}$ and compute its QR decomposition

$$C=\begin{array}{c}\begin{array}{cc} \scriptstyle m & \scriptstyle n-m\end{array}\\ \begin{bmatrix} S_2 & S_1\end{bmatrix}\end{array}\times\begin{array}{cc} & \scriptstyle m\\ \begin{array}{c}\scriptstyle m\\ \scriptstyle n-m\end{array} & \begin{bmatrix} R\\ 0\end{bmatrix}\end{array}\equiv S_2R; \tag{3.68}$$

3. Define $b=\zeta^2 R^{\mathrm{T}}a$;
4. Take $A_{12}=g_0a^{\mathrm{T}}$, $A_{22}=\eta I_m$ with $\eta=(g_0^{\mathrm{T}}H^{-1}g_0)/\zeta^2$;
5. Set $A=S\begin{bmatrix} H & A_{12}\\ A_{12}^{\mathrm{T}} & A_{22}\end{bmatrix}S^{\mathrm{T}}$, where $S=[S_1,S_2]$.

Note that by the construction, the matrix A is positive semidefinite when H is positive definite. This is because the Schur complement of H in the matrix $\begin{bmatrix} H & A_{12}\\ A_{12}^{\mathrm{T}} & A_{22}\end{bmatrix}$:

$$\begin{aligned}A_{22}-A_{12}^{\mathrm{T}}H^{-1}A_{12}&=A_{22}-ag_0^{\mathrm{T}}H^{-1}g_0a^{\mathrm{T}}=A_{22}-(g_0^{\mathrm{T}}H^{-1}g_0)aa^{\mathrm{T}}\\&=\eta I-(g_0^{\mathrm{T}}H^{-1}g_0)aa^{\mathrm{T}}=(g_0^{\mathrm{T}}H^{-1}g_0)(\|a\|^2I-aa^{\mathrm{T}})\end{aligned}$$

is positive semidefinite since H is positive definite and $g_0^{\mathrm{T}}H^{-1}g_0>0$.

Now we verify that CRQopt (1.1) with A, C, b constructed from the process above will yield pLGopt (2.23) with matrices H and g_0 and scalar $\gamma=\sqrt{1-\zeta^2}$, as desired.

Recall the definitions in (2.21):

$$g_0=S_1^{\mathrm{T}}b_0,\quad H=S_1^{\mathrm{T}}PAPS_1=S_1^{\mathrm{T}}AS_1\in\mathbb{R}^{(n-m)\times(n-m)}. \tag{3.69}$$

By the construction of A, $S_1^{\mathrm{T}}AS_1=H$, which is consistent with H defined in (3.69). Further recall that P is a projection matrix onto $\mathcal{N}(C^{\mathrm{T}})$ and the columns of S_1 form an orthonormal basis of $\mathcal{N}(C^{\mathrm{T}})$. So $P=S_1S_1^{\mathrm{T}}$. In addition, by the QR factorization (3.68), $(C^{\mathrm{T}})^{\dagger}=S_2R^{-\mathrm{T}}$, and so $n_0=(C^{\mathrm{T}})^{\dagger}b=S_2R^{-\mathrm{T}}b$. By the definition of matrix A, $S_1^{\mathrm{T}}AS_2=A_{12}$, we have

$$S_1^{\mathrm{T}}b_0=S_1^{\mathrm{T}}PAn_0=S_1^{\mathrm{T}}S_1S_1^{\mathrm{T}}AS_2R^{-\mathrm{T}}b=S_1^{\mathrm{T}}AS_2R^{-\mathrm{T}}b=\zeta^2A_{12}a=\zeta^2g_0a^{\mathrm{T}}a=g_0, \tag{3.70}$$

which is consistent with g_0 defined in (3.69). Finally,

$$\gamma=\sqrt{1-\|n_0\|^2}=\sqrt{1-\|S_2R^{-\mathrm{T}}b\|^2}=\sqrt{1-\|R^{-\mathrm{T}}b\|^2}=\sqrt{1-\|\zeta^2a\|^2}=\sqrt{1-\zeta^2}.$$

3.6.2 Numerical results

For testing purpose, we compute a solution v_* by the direct method [10] as a reference (exact) solution. We also compute $\kappa=\frac{\lambda_{\max}(H)-\lambda_*}{\lambda_{\min}(H)-\lambda_*}$ to examine the error bounds in Theorem 3.4.

The Lanczos algorithm (Algorithm 4) is applied to solve CRQopt (1.1) via QEPmin (2.18) and via LGopt (2.13). For each computed $v^{(k)}$, the kth iteration, we compute relative errors

$$\mathrm{err}_1=\frac{|(v^{(k)})^{\mathrm{T}}Av^{(k)}-v_*^{\mathrm{T}}Av_*|}{|v_*^{\mathrm{T}}Av_*|},\quad \mathrm{err}_2=\|v^{(k)}-v_*\|\quad\text{and}\quad \mathrm{err}_3=\frac{|\mu^{(k)}-\lambda_*|}{|\lambda_*|}.$$

Since $\|v_*\|=1$, the absolute error err$_2$ is also relative. The stoping criterion for solving QEPmin (2.18) is either $\delta_k^{\mathrm{QEPmin}}<10^{-15}$ or the number of Lanczos steps reaches $\mathtt{maxit}=200$, where $\delta_k^{\mathrm{QEPmin}}$ is defined in (3.27). The stoping criterion for solving LGopt (2.13) is either $\mathrm{NRes}_k^{\mathrm{LGopt}}<10^{-15}$ or the number of Lanczos steps reaches $\mathtt{maxit}=200$.

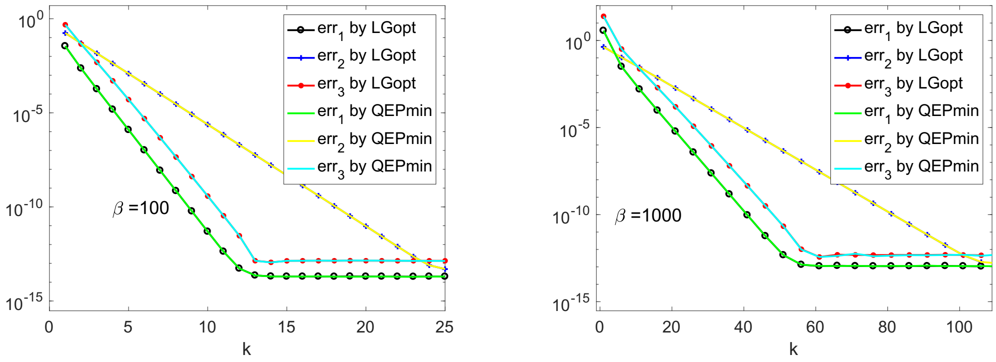

Figure 2: Example 3.2: history of err_1, err_2 and err_3 for the cases where $\beta=100$ (left) and $\beta=1000$ (right).

Example 3.2. In this example, we test the correctness and convergence behavior of the Lanczos algorithm to solve CRQopt (1.1). Let $n=1100$, $m=100$, $\alpha=1$, $\beta=100$ or 1000, and construct H as in (3.66) and g_0 as in (3.67). For (A,C,b), let $\zeta=0.9$ and a be random vector normalized to have norm $1/\zeta$ and then the rest follows Subsection 3.6.1 in constructing A, C and b.

The convergence histories for err_1, err_2 and err_3 are plotted in Fig. 2. It can be seen that all converge to the machine precision. err_1, err_2 and err_3 are the same, respectively, at every iteration whether CRQopt (1.1) is solved via QEPmin (2.18) or LGopt (2.13), which is consistent with our theory that solving rLGopt (3.5) is equivalent to solving rQEPmin (3.22).

Example 3.3. We illustrate the sharpness of the error bounds (3.49) in Theorem 3.4 and the relationship between the convergence rate of our Lanczos algorithm and κ.

The same test matrices as in Example 3.2, with $\beta=100$ and 1000 are used. We solve CRQopt (1.1) by solving QEPmin (2.18) and choose the same parameters as in Example 3.2. For $\alpha=1$ and $\beta=100$, We calculate

$$(\lambda_*,\kappa)=\begin{cases}(-42.6007,3.2706) & \text{for } (\alpha,\beta)=(1,100);\\ (-18.2629,52.8613) & \text{for } (\alpha,\beta)=(1,1000).\end{cases}$$

Judging from the corresponding κ, we expect our Lanczos algorithm will converge faster for the case $\beta=100$ than the case $\beta=1000$. We plot in Fig. 3 the convergence histories for

err_1 and its upper bound $\frac{16\gamma^2\|H-\lambda_*I\|}{v_*^{\mathrm{T}}Av_*}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}$ by (3.49a),

err_2 and its upper bound $4\gamma\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}$ by (3.49b),

err_3 and its upper bound $\frac{16}{|\lambda_*|}\|H-\lambda_*I\|\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}+\frac{4}{\gamma|\lambda_*|}\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}$ by (3.49c).

The bounds for err_1 and err_2 by (3.49a) and (3.49b) for both $\beta=100$ and $\beta=1000$ appear sharp. However, the bound for err_3 by (3.49c) is pessimistic. In the plots, err_3 goes to 0 at

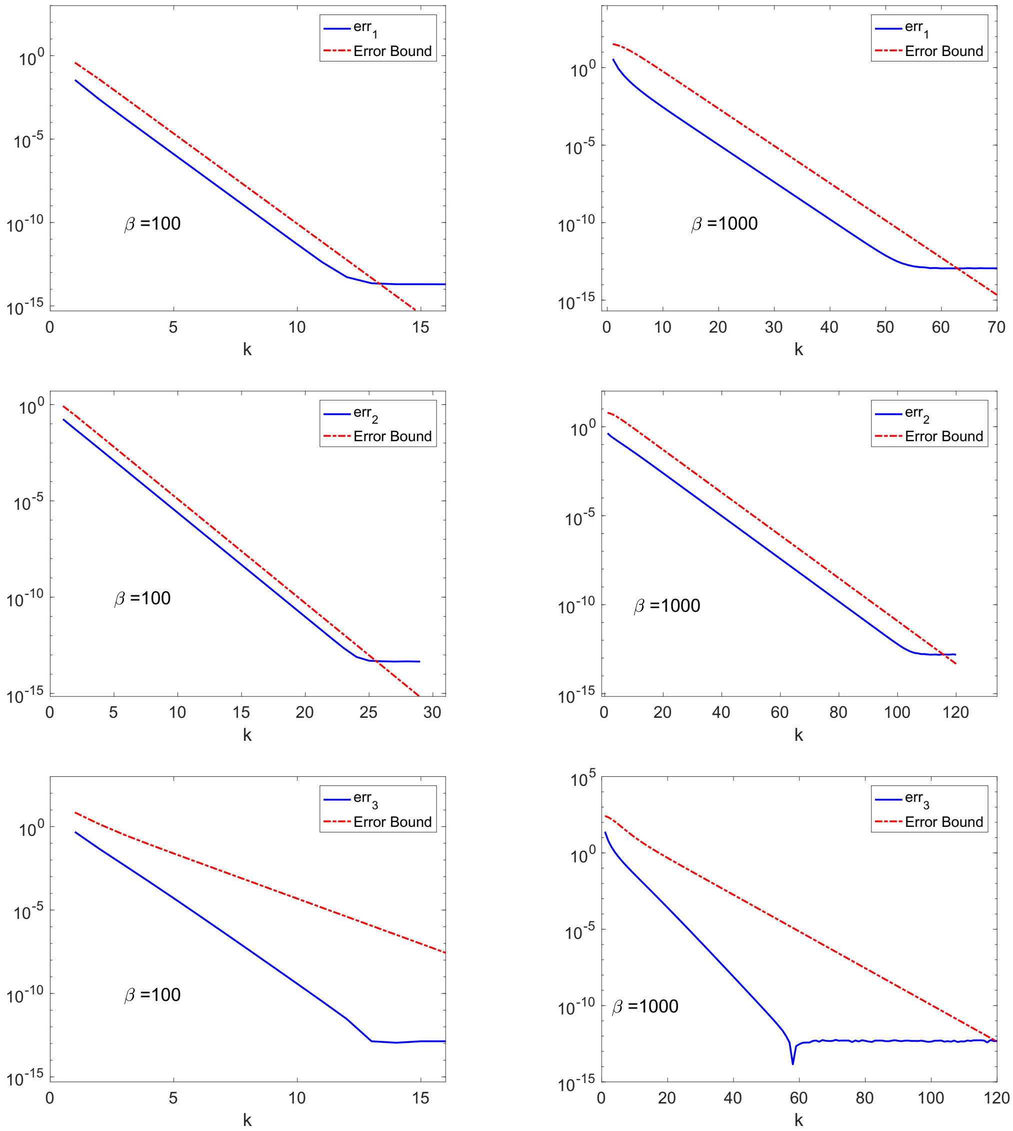

Figure 3: Example 3.3: histories for err$_1$ (first row), err$_2$ (second row), err$_3$ (third row) and their upper bounds for $\beta=100$ (left column) and $\beta=1000$ (right column).

about a similar rate of err$_1$, but the bounds by (3.49b) and (3.49c) for err$_3$ progress at the same rate as the bound by (3.49a) for err$_2$. As discussed in Remark 3.4, we unsuccessfully tried to establish a better bound for err$_3$, and are only able to offer an explanation.

As expected, err$_1$, err$_2$ and err$_3$ go to 0 faster for the case $\beta=100$ than the case $\beta=1000$. It is consistent with the convergence results in Theorem 3.4 that our Lanczos algorithm for CRQopt (1.1) converges faster when κ is smaller.

Example 3.4. We consider an example where the error bounds in Theorem 3.4 are pessimistic, while those by (3.60) can correctly reveal the speed of convergence. This occurs when CRQopt is a *"nearly the hard case"*, i.e., where the optimal value of the corresponding pLGopt (2.23) $\lambda_* \approx \lambda_{\min}(H)$. Specifically, we choose $n=1100$, $m=100$, $\zeta=0.9$, a a random vector with the norm $1/\zeta$, and $H=\operatorname{diag}(\tau^{\rm trans}_{0n-m-2}, \tau^{\rm trans}_{1n-m-2}, \cdots, \tau^{\rm trans}_{n-m-2n-m-2}, 1)$ with $(\alpha,\beta)=(2,1000)$ in (3.64) and (3.65), and $g_0=\left[e^{\eta}, e^{2\eta}, \cdots, e^{(n-m)\eta}\right]^{\rm T}$, where $\eta=-5\times 10^{-3}$. In this case, $\lambda_{\min}(H)=1$ and $\lambda_*=0.9845$, $\lambda_{\min}(H)\approx\lambda_*$ and $\kappa=\frac{\lambda_{\max}(H)-\lambda_*}{\lambda_{\min}(H)-\lambda_*}\approx 6.4466\times 10^4$. Thus it is a nearly the hard case and κ is large. We solve the associated CRQopt (1.1) via QEPmin (2.18). In Fig. 4, we plot the convergence history:

err$_1$, its upper bounds $\frac{16\gamma^2}{v_*^{\rm T}Av_*}\|H-\lambda_* I\|_2\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}$ by (3.49a), and

$$\frac{16\gamma^2}{v_*^{\rm T}Av_*}\|H-\lambda_* I\|_2\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)^2\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-2} \text{ by (3.60a)},$$

err$_2$, its upper bounds $4\gamma\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}$ by (3.49b), and

$$4\gamma\sqrt{\kappa}\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1} \text{ by (3.60b)},$$

err$_3$, its upper bounds $\frac{16}{|\lambda_*|}\|H-\lambda_* I\|_2\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}+\frac{4}{\gamma|\lambda_*|}\|b_0\|_2\sqrt{\kappa}\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}$ by (3.49c),

$$\text{and } \frac{16}{|\lambda_*|}\|H-\lambda_* I\|_2\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)^2\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-2}$$
$$+\frac{4}{\gamma|\lambda_*|}\|b_0\|_2\sqrt{\kappa}\left(\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}\right)\left[\Gamma_{\kappa_+}^{(k-1)}+\Gamma_{\kappa_+}^{-(k-1)}\right]^{-1} \text{ by (3.60c)}.$$

It can be observed that The error bounds by Theorem 3.4 decay much slower than err$_1$, err$_2$ and err$_3$ in this "near hard case". This is an example for which κ is large but κ_+ is small:

$$\kappa_+ := \frac{\theta_{\max}-\lambda_*}{\theta_2-\lambda_*}\approx 983.7702,$$

As commented in Remark 3.3, sharper bounds (3.60) should be used. We can see that the bounds (3.60) correctly reflect the slope of the convergence, but they are still larger than the actual errors by several order of magnitudes. This is due to the fact that in the proof of the bounds (3.60), we use $\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-1}$ and $\left[\Gamma_\kappa^k+\Gamma_\kappa^{-k}\right]^{-2}$ to reflect the convergence trend. We select polynomials such that $\max_{1\leqslant i\leqslant n-m}|\psi_k(\theta_i-\lambda_*)|=\max_{2\leqslant i\leqslant n-m}|\psi_k(\theta_i-\lambda_*)|$ by setting $\psi_k(\theta_1-\lambda_*)=0$. In this case the coefficients involving $\frac{\theta_{\max}-\theta_{\min}}{\theta_{\min}-\lambda_*}$ is large in nearly the hard case when $\theta_{\min}\approx\lambda_*$.

Example 3.5. In this example, we test the effectiveness of the residual bound $\delta_k^{\rm QEPmin}$ in (3.27). We use the same test problem as in Example 3.2 for both $\beta=100$ and $\beta=1000$. We run our Lanczos algorithm for QEPmin (2.18) and record the residual ${\rm NRes}_k^{\rm QEPmin}$ and its bound $\delta_k^{\rm QEPmin}$ defined in (3.27) for every Lanczos step in Fig. 5. We observe that both ${\rm NRes}_k^{\rm QEPmin}$ and $\delta_k^{\rm QEPmin}$ in (3.27) converge at the same rate, suggesting $\delta_k^{\rm QEPmin}$ is an effective upper bound of the residual ${\rm NRes}_k^{\rm QEPmin}$.

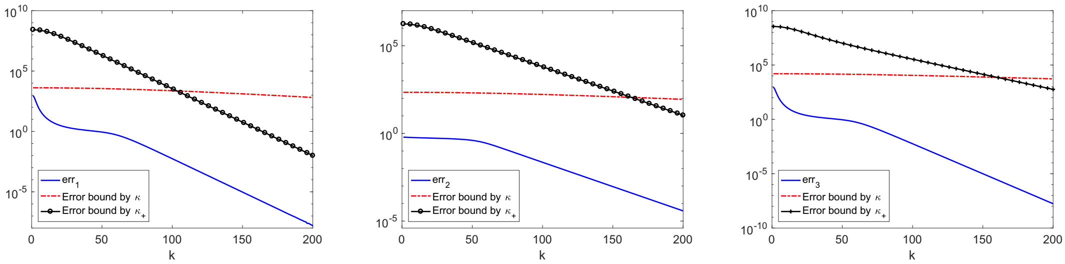

Figure 4: Example 3.4: histories of err_1, err_2, err_3 and their upper bounds. "Error bound by κ" and "Error bound by κ_+" means upper bounds in (3.49) and (3.60), respectively.

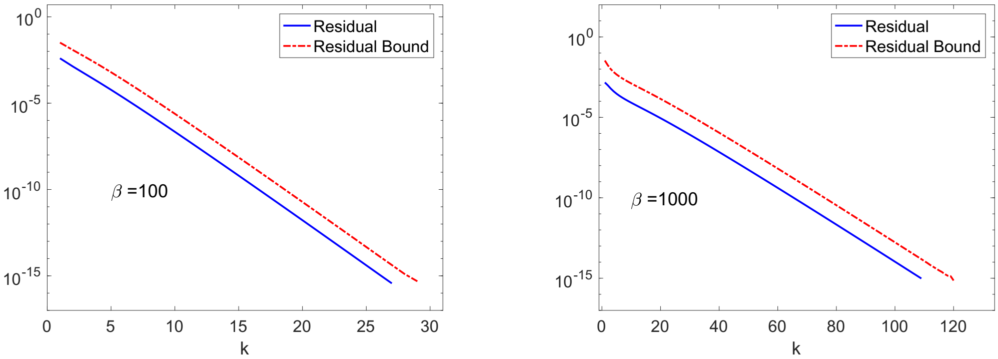

Figure 5: Example 3.5: relative residual of QEP $\mathrm{NRes}_k^{\mathrm{QEPmin}}$ and the bound of the relative residual $\delta_k^{\mathrm{QEPmin}}$ for the case where $\beta = 100$ (left) and $\beta = 1000$ (right).

4 Application to the constrained clustering

In this section, we use semi-supervised learning for clustering as an application of CRQopt (1.1). We first discuss unconstrained clustering in Section 4.1 and then discuss a new model for constrained clustering in Section 4.2. Numerical experiments are shown in Section 4.3.

4.1 Unconstrained clustering

Clustering is an important technique for data analysis and is widely used in machine learning [8, Chapter 14.5.3], bioinformatics [32], social science [26] and image analysis [36]. Clustering uses some similarity metric to group data into different categories. In this section, we discuss the normalized cut, a spectral clustering method that are popular for image segmentation [36,39].

Given an undirected graph $G = (\mathcal{V},\mathcal{E})$ whose edge weights are represented by an affinity matrix $W = [w_{ij}]$, we define the *cut* of a partition on its vertices $\mathcal{V}$ into two disjoint sets

$\mathcal{A}$ and $\mathcal{B}$, i.e., $\mathcal{A}\cup\mathcal{B}=\mathcal{V}$, $\mathcal{A}\cap\mathcal{B}=\varnothing$ as

$$\text{cut}(\mathcal{A},\mathcal{B})=\sum_{i\in\mathcal{A},j\in\mathcal{B}} w_{ij}. \tag{4.1}$$

Intuitively one would minimize the cut to achieve an optimal bipartition of the graph G, but it often results in a partition $(\mathcal{A},\mathcal{B})$ with one of them containing only a few isolated vortices in the graph while the other containing the rest. Such a bipartition is not balanced and not useful in practice. To avoid such an unnatural bias that leads to small sets of isolated vortices, the following *normalized cut* [36] is introduced:

$$\text{Ncut}(\mathcal{A},\mathcal{B})=\frac{\text{cut}(\mathcal{A},\mathcal{B})}{\text{vol}(\mathcal{A})}+\frac{\text{cut}(\mathcal{A},\mathcal{B})}{\text{vol}(\mathcal{B})}, \tag{4.2}$$

where

$$\text{vol}(\mathcal{A})=\sum_{i\in\mathcal{A},j\in\mathcal{V}} w_{ij} \quad \text{and} \quad \text{vol}(\mathcal{B})=\sum_{i\in\mathcal{B},j\in\mathcal{V}} w_{ij}.$$

It turns out that minimizing $\text{Ncut}(\mathcal{A},\mathcal{B})$ usually yields a more balanced bipartition. Let

$$c_+=\sqrt{\frac{\text{vol}(\mathcal{B})}{\text{vol}(\mathcal{A})\cdot\text{vol}(\mathcal{V})}} \quad \text{and} \quad c_-=-\sqrt{\frac{\text{vol}(\mathcal{A})}{\text{vol}(\mathcal{B})\cdot\text{vol}(\mathcal{V})}},$$

and $x\in\mathbb{R}^n$ ($n=|\mathcal{V}|$, the cardinality of $\mathcal{V}$) be the indicator vector for bipartition $(\mathcal{A},\mathcal{B})$, i.e.,

$$x_{(i)}=\begin{cases} c_+, & i\in\mathcal{A}, \\ c_-, & i\in\mathcal{B}, \end{cases} \tag{4.3}$$

and D be a diagonal matrix with the row sums of W on the diagonal, i.e., $D=\text{diag}(W\mathbf{1})$. Then it can be verified that

$$\text{Ncut}(\mathcal{A},\mathcal{B})=x^{\mathrm{T}}(D-W)x, \quad x^{\mathrm{T}}Dx=1, \quad (Dx)^{\mathrm{T}}\mathbf{1}=0,$$

where $\mathbf{1}$ is a vector of ones. Therefore in order to minimize $\text{Ncut}(\mathcal{A},\mathcal{B})$, we will solve the following combinatorial optimization problem

$$\begin{cases} \min\ x^{\mathrm{T}}(D-W)x, & (4.4\text{a}) \\ \quad \text{s.t. } x_{(i)}\in\{c_+,c_-\}, & (4.4\text{b}) \\ \qquad x^{\mathrm{T}}Dx=1, & (4.4\text{c}) \\ \qquad (Dx)^{\mathrm{T}}\mathbf{1}=0. & (4.4\text{d}) \end{cases}$$

However, the problem (4.4) is a discrete optimization problem and known to be NP-complete. A common practice to make it numerical feasible is to relax x to a real vector and solve instead the following optimization problem

$$\begin{cases} \min\ x^{\mathrm{T}}(D-W)x, & (4.5\text{a})\\ \ \text{s.t.}\ x^{\mathrm{T}}Dx=1, & (4.5\text{b})\\ \qquad (Dx)^{\mathrm{T}}\mathbf{1}=0, & (4.5\text{c})\\ \qquad x\in\mathbb{R}^n. & (4.5\text{d}) \end{cases}$$

Under the assumption that D is positive definite, by the Courant-Fisher variational principle [13, Sec 8.1.1], solving (4.5) is equivalent to finding the eigenvector x corresponding to the second smallest eigenvalue of the generalized symmetric definite eigenproblem

$$(D-W)x=\lambda Dx.$$

Note that the setting here is different from the one in [36], where the indicator vector $x_{(i)}\in\{1,-b\}$ and $b=\frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{B})}$. Instead of minimizing a quotient of two quadratic functions in [36], we use the constraint that $x^{\mathrm{T}}Dx=1$. The model (4.4) is similar to the one in [39, section 5.1], where they use the number of vertices in the sets $\mathcal{A}$ and $\mathcal{B}$ instead of the volumes. The model (4.4) is derived in a similar way to the derivation in [39, section 5.1].

4.2 Constrained clustering

When partial grouping information is known in advance, we can use partial grouping information to set up different models for better clustering. These models are known as constrained clustering. Existing methods for constrained spectral clustering includes implicitly incorporating the constraints into Laplacians [3,18] and imposing the constraints in linear forms [6,41,42] or bilinear forms [40].

We encode the partial grouping information into linear constraints, which can be either homogeneous [42] or nonhomogeneous [6,41]. In [6], the authors set up a model where the objective function is the quotient of two quadratic functions and used hard coding for the known associations of pixels to specific classes in terms of linear constraints. In [41], the authors used a model for which the objective function is quadratic and encoded known labels by linear constraints. This is an approach that we take to set up the model.

Let $\mathcal{I}=\{i_1,\cdots,i_\ell\}$ be the index set for which we have the prior information such as $\mathcal{I}\subseteq\mathcal{A}$. According to (4.3), we set $x_{(i)}=c_+$ for $i\in\mathcal{I}$. Similarly, let $\mathcal{J}=\{j_1,\cdots,j_k\}$ be the index set for which we have the prior information that $\mathcal{J}\subseteq\mathcal{B}$, and we set $x_{(j)}=c_-$ for $j\in\mathcal{J}$. This leads to the following discrete constrained normalized cut problem

$$\begin{cases} \min\ x^{\mathrm{T}}(D-W)x, & \text{(4.6a)}\\ \quad\text{s.t. } x_{(i)}\in\{c_+,c_-\}, & \text{(4.6b)}\\ \qquad x^{\mathrm{T}}Dx=1, & \text{(4.6c)}\\ \qquad (Dx)^{\mathrm{T}}\mathbf{1}=0, & \text{(4.6d)}\\ \qquad x_{(i)}=c_+ \quad \text{for } i\in\mathcal{I}, & \text{(4.6e)}\\ \qquad x_{(i)}=c_- \quad \text{for } i\in\mathcal{J}. & \text{(4.6f)} \end{cases}$$

However, there are two imminent issues associated with the model (4.6):

1. the combinatorial optimization (4.6) is NP-hard;
2. the model is incomplete because to calculate c_+ and c_- we need to know $\mathrm{vol}(\mathcal{A})$ and $\mathrm{vol}(\mathcal{B})$, which are unknown before the clustering.

Common workarounds, which we use, are as follows. For the first issue, we relax the model (4.6) by allowing x to be a real vector, i.e., $x\in\mathbb{R}^n$. For the second issue, we use $\frac{\mathrm{vol}(\mathcal{J})}{\mathrm{vol}(\mathcal{I})}$ as an estimate of $\frac{\mathrm{vol}(\mathcal{B})}{\mathrm{vol}(\mathcal{A})}$ to get

$$c_+\approx\hat{c}_+=\sqrt{\frac{\mathrm{vol}(\mathcal{J})}{\mathrm{vol}(\mathcal{I})\cdot\mathrm{vol}(\mathcal{V})}},\quad c_-\approx\hat{c}_-=-\sqrt{\frac{\mathrm{vol}(\mathcal{I})}{\mathrm{vol}(\mathcal{J})\cdot\mathrm{vol}(\mathcal{V})}}.$$

By these relaxation, we reach a computational feasible model:

$$\begin{cases} \min\ x^{\mathrm{T}}(D-W)x, & \text{(4.7a)}\\ \quad\text{s.t. } x^{\mathrm{T}}Dx=1, & \text{(4.7b)}\\ \qquad (Dx)^{\mathrm{T}}\mathbf{1}=0, & \text{(4.7c)}\\ \qquad x_{(i)}=\hat{c}_+, \quad i\in\mathcal{I}, & \text{(4.7d)}\\ \qquad x_{(i)}=\hat{c}_-. \quad i\in\mathcal{J}, & \text{(4.7e)} \end{cases}$$

The last three equations are linear constraints and can be collectively written as a linear system of equations:

$$N^{\mathrm{T}}x=b.$$

Let $v=D^{1/2}x$, and define

$$A=D^{-1/2}(D-W)D^{-1/2} \quad\text{and}\quad C=D^{-1/2}N.$$

Then the optimization problem (4.7) is turned into CRQopt (1.1) with matrices A, C and b just defined.

4.3 Numerical results

In this section, we show the numerical results of the constrained clustering for the segmentation of a set of images listed in Table 1 and Fig. 6. All experiments were conducted on a PC with Intel Core i7-4770K CPU@3.5GHz and 16-GB RAM. CRQopt (1.1) is solved via solving QEPmin (2.18). The minimum and maximum numbers of Lanczos steps are `minit`$=120$ and `maxit`$=300$, respectively. The stopping criterion is $\delta_k^{\rm QEPmin}<8\times 10^{-5}$. We check the stopping criterion every five iterations.

For a grayscale image, we can construct a weighted graph $G=(\mathcal{V},\mathcal{E})$ by taking each pixel as a node and connecting each pair (i,j) of pixel i and j by an edge with a weight given by

$$w_{ij}=e^{-\frac{\|F(i)-F(j)\|_2^2}{\delta_F}}\times\begin{cases}1 & \text{if } \|X(i)-X(j)\|_\infty<r,\\ 0 & \text{otherwise}\,,\end{cases} \tag{4.8}$$

where δ_F and r are chosen parameters, F is the brightness value and X is the location of a pixel [36].[‖] We take $\delta_F=\delta\cdot\max_{i,j}\|F(i)-F(j)\|_2^2$ for some parameter δ to be specified below. The definition of weight in (4.8) ensures that every pixel is connected with an edge to at most $(2r+1)^2$ other pixels.

Table 1 lists the values of key parameters used in our experiments. r is taken either 5 or 10, and thus the weight matrix W is sparse, which in turn makes the matrix A in CRQopt (1.1) sparse, too. Note that for the Crab image, the contrast between the upper right of the object and the background is not significant. Therefore, r is chosen to be twice as much as other images to ensure the weight matrix correctly reflect the connectivity of the graph. δ is around 0.1, to be consistent with the statement in [36] that "δ_F is typically set to 10 to 20 percent of the total range of the feature distance function". Finally, the number m of linear constraints is small yielding CRQopt (1.1) with $m\ll n$.

[‖] In a 2-D image, pixel i may naturally be represented by (i_x,i_y) where i_x and i_y are two integers.

Table 1: The number of pixels n, parameters δ and r and size m of linear constraints.

Image	Number of pixels n	δ	r	m
Flower	30,000	0.1	5	24
Road	50,268	0.1	5	46
Crab	143,000	0.1	10	32
Camel	240,057	0.08	5	24
Dog	395,520	0.1	5	33
Face1	562,500	0.1	5	31
Face2	922,560	0.1	5	19
Daisy	1,024,000	0.08	5	29
Daisy2	1,024,000	0.08	5	59

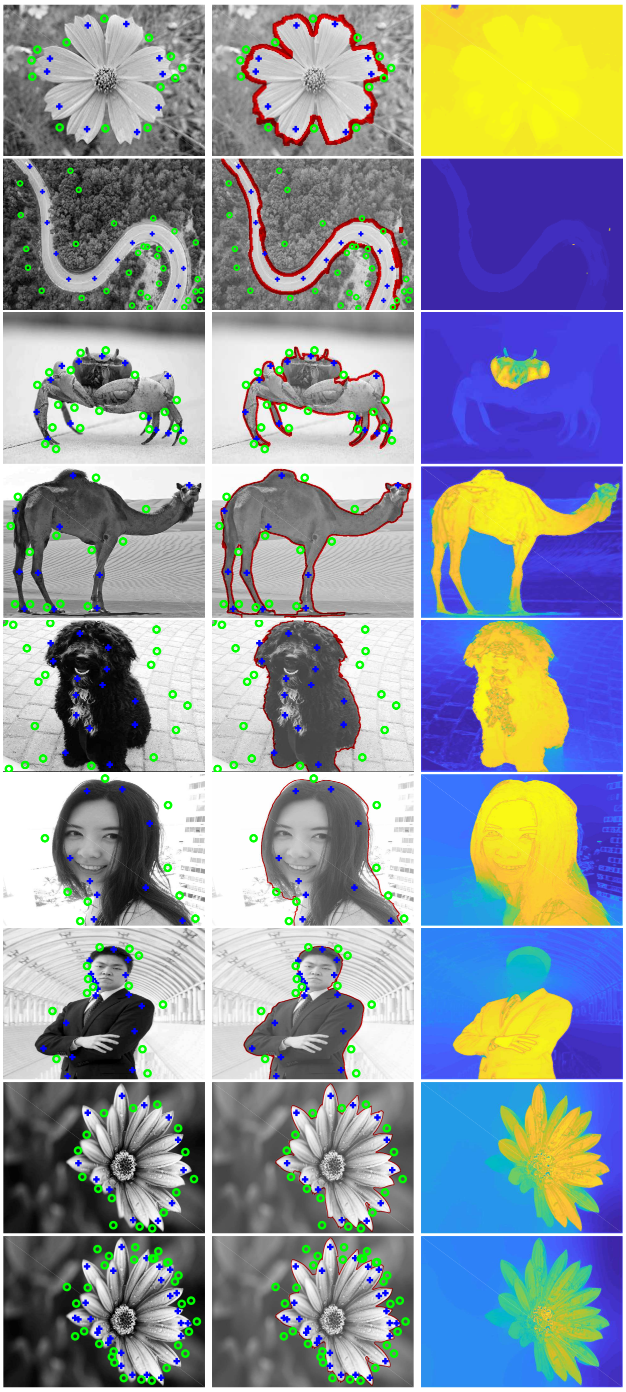

Figure 6: The left, middle and right columns are labels, results of image cut and the heat maps of the solutions by the Lanczos algorithm for CRQopt, respectively. Images from top to bottom are Flower, Road, Crab, Camel, Dog, Face1, Face2, Daisy and Daisy2, respectively.

Table 2: Runtime (in seconds) and number of Lanczos steps.

Image	Run Time	Lanczos steps
Flower	4.61	210
Road	14.92	200
Crab	21.58	135
Camel	31.12	300
Dog	22.33	135
Face1	67.46	215
Face2	35.54	165
Daisy	84.09	235
Daisy2	105.80	245

Table 3: Runtime for Fast-GE-2.0, projected power method and the Lanczos algorithm.

Image	Fast-GE-2.0	Projected Power Method	Lanczos algorithm
Crab	47.13 s	446.76 s	21.58 s
Daisy	1572.81 s	3+ hours	84.09 s
Daisy2	1319.58 s	3+ hours	105.80 s

Fig. 6 shows that the results of the model (4.7) for the image segmentation indeed agree with natural visual separation of the object and the background. Table 2 displays the wall-clock runtime and the numbers of Lanczos steps used for the images.

We note that Daisy and Daisy2 are the same image but with two different ways of prior partial labeling. For both ways of prior partial labelling, the computed image cuts look equally well. The purpose of experiments on Daisy and Daisy2 is to observe how the size m of the linear constraints may affect running time. Daisy has 29 linear constraints while Daisy2 has 59. As shown in Table 2, the Lanczos algorithm took 84.09 seconds for Daisy and 105.80 seconds for Daisy2. It suggests that the larger m is, the more times the Lanczos algorithm takes to solve the associated CRQopt. This is because the matrix-vector product Px does more work as m increases.

In Table 3, we show the running time of Fast-GE-2.0 [18], the projected power method [41], and the Lanczos algorithm for selected images. For comparable segmentation quality, the runtime of the Lanczos algorithm is significantly less than Fast-GE-2.0 and the projected power method.

5 Conclusions

Although the constrained Rayleigh quotient optimization problem (CRQopt) (1.1), also known as the constrained eigenvalue problem, has been around since 1970s, some of the mathematical claims were not rigorously justified. There are very few numerical methods that are suitable for large scale CRQopt (1.1), such as those arising from constrained image segmentation. The projected power method [41] converges too slow while the method in [14] is for the homogeneous constraints only. Eigenvalue optimization method [6] could be too expensive. In this paper, we conducted a systematical and rigorous theoretical study of the problem and, as a result, devised an efficient Lanczos algorithm for large scale CRQopt (1.1). We perform a detailed convergence analysis of the Lanczos algorithm. As an application, we apply our Lanczos algorithm to the image cut problem with partial prior labeling. Numerical experiments demonstrate the effectiveness of the algorithm in terms of accuracy and superior efficiency compared to Fast-GE-2.0 [18] and the projected power method [41]. Future work include the treatment of rLGopt (3.5) in nearly the hard case and applications of Lanczos algorithms on other machine learning problems such as outlier removal [25], semi-supervised kernel PCA [31], and transductive learning [19].

Although our presentation in this paper is restricted to the real numbers, their extensions to the complex version of CRQopt (1.1)

$$\min_{v\in\mathbb{C}^n} v^{\rm H}Av \quad \text{s.t.} \quad v^{\rm H}v=1 \quad \text{and} \quad C^{\rm H}v=b$$

are rather straightforward, where $A\in\mathbb{C}^{n\times n}$ is Hermitian, i.e., $A=A^{\rm H}$, $C\in\mathbb{C}^{n\times m}$. Essentially, all we need to do is to replace all transposes $\cdot^{\rm T}$ by complex conjugate transposes $\cdot^{\rm H}$. We also note that we can also easily extend the discussion of this paper to the model

$$\min_{x\in\mathbb{R}^n} x^{\rm T}Ax \quad \text{s.t.} \quad x^{\rm T}Bx=1 \quad \text{and} \quad C^{\rm T}x=b, \tag{5.1}$$

where B is a symmetric positive definite matrix. In fact, let $B=LL^{\rm T}$ be the Cholesky decomposition of B and $v=Lx$, then (5.1) is transformed to CRQopt (1.1) with $A:=L^{-{\rm T}}AL$ and $C:=L^{-{\rm T}}C$.

Acknowledgments

We would like to thank Mr. Ning Wan for sharing his study notes on CRQopt, Dr. Yanwen Luo for his help in the proof for Lemma 2.8, Dr. Chengming Jiang for providing his FAST-GE2.0 implementation of the constrained image segmentation, and Mr. Michael Ragone for his comments on an early version of this manuscript. Y. Zhou and Z. Bai were supported in part by NSF grants DMS-1522697, CCF-1527091 and DMS-1913364. R. Li was supported in part by NSF grants CCF-1527104 and DMS-1719620. We are grateful for the anonymous referees for their careful review of the manuscript and constructive suggestions.

Appendices

A Solve secular equation

We are interested in computing the smallest root λ_* of the secular function

$$\chi(\lambda):=\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda-\theta_i)^2}-\gamma^2, \tag{A.1}$$

where it is assumed

$$\gamma>0,\quad \theta_1\leqslant\theta_2\leqslant\cdots\leqslant\theta_n \text{ and either } \xi_1\neq 0 \text{ or } \xi_1=0 \text{ but } \lim\nolimits_{\lambda\to\theta_1^-}\chi(\lambda)>0.$$

Those assumptions guarantee that $\chi(\lambda)$ has a unique zero λ_* in $(-\infty,\theta_1)$. This is due to the facts

$$\lim_{\lambda\to-\infty}\chi(\lambda)=-\gamma^2<0,\quad \lim_{\lambda\to\theta_1^-}\chi(\lambda)>0,\quad \text{and}\quad \chi'(\lambda)=-2\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda-\theta_i)^3}>0 \text{ for } \lambda<\theta_1.$$

First, we find an initial lower bound $\alpha^{(0)}$ of λ_*, i.e., $\alpha^{(0)}<\theta_1$ such that $\chi(\alpha^{(0)})<0$. Note

$$\chi(\lambda)\leqslant\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda-\theta_1)^2}-\gamma^2 \quad \text{for } \lambda<\theta_1.$$

One such $\alpha^{(0)}$ can be found by solving

$$\sum_{i=1}^{n}\frac{\xi_i^2}{(\alpha^{(0)}-\theta_1)^2}-\gamma^2=0 \quad\Rightarrow\quad \alpha^{(0)}=\theta_1-\delta_0 \text{ with } \delta_0=\frac{1}{\gamma}\sqrt{\sum_{i=1}^{n}\xi_i^2}.$$

We conclude that $\lambda_*\in[\alpha^{(0)},\beta^{(0)}]$, where $\beta^{(0)}=\theta_1$. Quantities $\alpha^{(k)}$ and $\beta^{(k)}$ will be determined during our iterative process to be described such that $\lambda_*\in[\alpha^{(k)},\beta^{(k)}]$.

Without loss of generality, we may assume that

$$\text{if } \theta_1=\cdots=\theta_d<\theta_{d+1}, \text{ then } \xi_2=\cdots=\xi_d=0.$$

Let

$$j_0=\min\{i:\ \xi_i\neq 0\}. \tag{A.2}$$

To find the initial guess of the root, we solve

$$\frac{\xi_{j_0}^2}{(\lambda-\theta_{j_0})^2}+\underbrace{\sum_{i=j_0+1}^{n}\frac{\xi_i^2}{([\theta_{j_0}-\delta_0]-\theta_i)^2}-\gamma^2}_{=:-\eta}=0$$

for λ to get

$$\lambda^{(0)}=\begin{cases}\theta_{j_0}-|\xi_{j_0}|/\sqrt{\eta}, & \text{if } \eta>0,\\ \theta_{j_0}-\delta_0/2, & \text{if } \eta\leqslant 0,\end{cases}$$

where the second case is based on bisection.

For the iterative scheme, suppose we have an approximation $\lambda^{(k)}\approx\lambda_*$. First, the interval $(\alpha^{(k)},\beta^{(k)})$ will be updated as

$$\alpha^{(k+1)}\leftarrow\lambda^{(k)} \text{ and } \beta^{(k+1)}\leftarrow\beta^{(k)} \text{ if } \chi(\lambda^{(k)})<0,$$
$$\beta^{(k+1)}\leftarrow\lambda^{(k)} \text{ and } \alpha^{(k+1)}\leftarrow\alpha^{(k)} \text{ if } \chi(\lambda^{(k)})>0.$$

Then we find the next approximation $\lambda^{(k+1)}$. For that purpose, we seek to approximate χ, in the neighborhood of $\lambda^{(k)}$, by

$$g(\lambda):=-b+\frac{a}{(\lambda-\theta_{j_0})^2}\approx\chi(\lambda),$$

such that

$$\begin{aligned} g(\lambda^{(k)})\equiv -b+\frac{a}{(\lambda^{(k)}-\theta_{j_0})^2} &=\chi(\lambda^{(k)})=\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda^{(k)}-\theta_i)^2}-\gamma^2,\\ g'(\lambda^{(k)})\equiv -2\frac{a}{(\lambda^{(k)}-\theta_{j_0})^3} &=\chi'(\lambda^{(k)})=-2\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda^{(k)}-\theta_i)^3},\end{aligned}$$

yielding

$$\begin{aligned} a&=-\frac{1}{2}(\lambda^{(k)}-\theta_{j_0})^3\chi'(\lambda^{(k)})=(\lambda^{(k)}-\theta_{j_0})^3\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda^{(k)}-\theta_i)^3}>0,\\ b&=\frac{a}{(\lambda^{(k)}-\theta_{j_0})^2}-\chi(\lambda^{(k)})=(\lambda^{(k)}-\theta_{j_0})\sum_{i=1}^{n}\frac{\xi_i^2}{(\lambda^{(k)}-\theta_i)^3}-\chi(\lambda^{(k)}).\end{aligned}$$

Ideally, $b>0$ so that $g(\lambda)=0$ has a solution in $(-\infty,\theta_{j_0})$. Assuming $b>0$, we find the next approximation $\lambda^{(k+1)}\approx\lambda_*$ is given by

$$\lambda^{(k+1)}=\theta_1-\sqrt{a/b}. \tag{A.3}$$

Now if $b\leqslant 0$ (then $\lambda^{(k+1)}$ as in (A.3) is undefined) or if $\lambda^{(k+1)}\notin(\alpha,\beta)$, we let $\lambda^{(k+1)}$ be $(\alpha^{(k+1)}+\beta^{(k+1)})/2$ according to bisection method.

B Proof of the equivalence between CRQopt and the eigenvalue optimization problem

Consider CRQopt (1.1), suppose $U\in\mathbb{R}^{n\times(n-m)}$ has full column rank and that $\mathcal{R}(U)=\mathcal{N}(C^{\mathrm{T}})$ and let $u\in\mathbb{R}^n$ satisfies $C^{\mathrm{T}}u=\sqrt{n}b$. Define

$$\widehat{C}=\begin{bmatrix}C^{\mathrm{T}} & -\sqrt{n}b\end{bmatrix},\quad N=\begin{array}{c c} & \begin{array}{cc} \scriptstyle n-m & \scriptstyle 1 \end{array} \\ \begin{array}{c}\scriptstyle n\\ \scriptstyle 1\end{array} & \begin{bmatrix} U & u\\ 0 & 1\end{bmatrix}\end{array} \tag{B.1}$$

and

$$L=N^{\mathrm{T}}\begin{bmatrix}A & 0\\ 0 & 0\end{bmatrix}N,\quad E=N^{\mathrm{T}}\begin{bmatrix}-\frac{I}{n+1} & 0\\ 0 & 1-\frac{1}{n+1}\end{bmatrix}N,\quad M=N^{\mathrm{T}}\begin{bmatrix}I_n & 0\\ 0 & 0\end{bmatrix}N.$$

Note that it is easy to see that $\mathcal{R}(N)=\mathcal{N}(\widehat{C})$.

In this appendix we prove that CRQopt (1.1) is equivalent to the following eigenvalue optimization problem

$$\max_{t\in\mathbb{R}}\lambda_{\min}(L+tE,M), \tag{B.2}$$

where $\lambda_{\min}(L+tE,M)$ is the smallest eigenvalue of $(L+tE)x=\lambda Mx$. This equivalency was initially established by Eriksson, Olsson and Kahl [6]. However, the statements presented here are stronger than the related ones in [6]. For examples, we will prove M is positive definite, and we can use 'max' in (B.2) instead of 'sup' in [6].

Let $\widetilde{v}=\sqrt{n}v$, $\widehat{v}=\begin{bmatrix}\widetilde{v}\\ 1\end{bmatrix}$, $\widehat{A}=\begin{bmatrix}A & 0\\ 0 & 0\end{bmatrix}$ and $\widehat{B}=\begin{bmatrix}I_n & 0\\ 0 & 0\end{bmatrix}$. Then v is a minimizer of CRQopt (1.1) if and only if $\widehat{v}$ is a minimizer of

$$\min\frac{\widehat{v}^{\mathrm{T}}\widehat{A}\widehat{v}}{\widehat{v}^{\mathrm{T}}\widehat{B}\widehat{v}},\quad \text{s.t.}\quad \widehat{v}^2_{(n+1)}=1,\quad \widehat{v}^{\mathrm{T}}\widehat{v}=n+1,\quad \widehat{C}\widehat{v}=0. \tag{B.3}$$

Since $\mathcal{R}(N)=\mathcal{N}(\widehat{C})$, for any $\widehat{v}$ satisfying $\widehat{C}\widehat{v}=0$, there exists $\widehat{y}\in\mathbb{R}^{n-m+1}$ such that $\widehat{v}=N\widehat{y}$, N is defined in (B.1). By the matrix structure in (B.1), we know that $\widehat{v}^2_{(n+1)}=1$ if and only if $\widehat{y}^2_{(n-m+1)}=1$. Therefore, solving (B.3) is equivalent to solving

$$\min\frac{\widehat{y}^{\mathrm{T}}L\widehat{y}}{\widehat{y}M\widehat{y}},\quad \text{s.t.}\quad \widehat{y}^2_{(n-m+1)}-1=0,\quad \widehat{y}^{\mathrm{T}}N^{\mathrm{T}}N\widehat{y}=n+1. \tag{B.4}$$

To prove (B.4) is equivalent to its dual problem, we use the following result on the duality of the quadratic constrained optimization problems.

Lemma B.1 ([6, Corollary 1]). *Let $y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2$ be a positive semidefinite quadratic form. If there exists y such that $y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3<0$ and if A_3 is positive semidefinite, then the primal problem*

$$\inf_{y}\frac{y^{\mathrm{T}}A_1y+2b_1^{\mathrm{T}}y+c_1}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2},\quad \text{s.t.}\quad y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3=0$$

and the dual problem

$$\sup_{\lambda}\inf_{y}\frac{y^{\mathrm{T}}(A_1+\lambda A_3)y+2(b_1+\lambda b_3)^{\mathrm{T}}y+(c_1+\lambda c_3)}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2}$$

has no duality gap.

Proof. See [6, Corollary 1]. □

With the help of Lemma B.1, we have the following theorem to show that there is no duality gap between the optimization problem (B.4) and its dual problem.

Theorem B.1 ([6, Theorem 1]). *Let* $\widehat{A}_i=\begin{bmatrix}A_i & b_i\\ b_i^{\mathrm{T}} & c_i\end{bmatrix}$ *for* $i=1,2,3$. *If* $\widehat{A}_2$ *and* A_3 *are positive semidefinite and if there exists* $\widehat{y}$ *such that* $\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}<n+1$ *and* $\widehat{y}_{n+1}^2=1$, *then the primal problem*

$$\inf_{y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3=n+1}\frac{y^{\mathrm{T}}A_1y+2b_1^{\mathrm{T}}y+c_1}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2}=\inf_{\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}=n+1,\widehat{y}_{n+1}^2=1}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}} \tag{B.5}$$

and its dual

$$\sup_{t}\inf_{\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}=n+1}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}+t\widehat{y}_{n+1}^2-t}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}}$$

has no duality gap.

Proof. Let γ_* be the optimal value of (B.5), then

$$\begin{aligned}
\gamma_*&=\inf_{\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}=n+1,\widehat{y}_{n+1}^2=1}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}}\\
&=\sup_{t}\inf_{\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}=n+1,\widehat{y}_{n+1}^2=1}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}+t\widehat{y}_{n+1}^2-t}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}}\\
&\geqslant\sup_{t}\inf_{\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}=n+1}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}+t\widehat{y}_{n+1}^2-t}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}}\\
&\geqslant\sup_{t,\lambda}\inf_{\widehat{y}}\frac{\widehat{y}^{\mathrm{T}}\widehat{A}_1\widehat{y}+t\widehat{y}_{n+1}^2-t+\lambda(\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}-(n+1))}{\widehat{y}^{\mathrm{T}}\widehat{A}_2\widehat{y}}\\
&=\sup_{t,\lambda}\inf_{\widehat{y}}\frac{y^{\mathrm{T}}A_1y+2b_1^{\mathrm{T}}y+c_1+t\widehat{y}_{n+1}^2-t+\lambda(y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3-(n+1))}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2}\\
&=\sup_{t,\lambda}\inf_{\widehat{y}_{n+1}^2=1}\frac{y^{\mathrm{T}}A_1y+2b_1^{\mathrm{T}}y+c_1+\lambda(y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3-(n+1))}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2} &\text{(B.6)}\\
&=\inf_{y^{\mathrm{T}}A_3y+2b_3^{\mathrm{T}}y+c_3=n+1}\frac{y^{\mathrm{T}}A_1y+2b_1^{\mathrm{T}}y+c_1}{y^{\mathrm{T}}A_2y+2b_2^{\mathrm{T}}y+c_2}=\gamma_*, &\text{(B.7)}
\end{aligned}$$

where (B.6) and (B.7) apply Lemma B.1. □

Remark B.1. One of the conditions in [6, Theorem 1] is "$\widehat{A}_3$ is positive semidefinite". However, the proof of Theorem B.1 applies Lemma B.1, which requires A_3 to be positive semidefinite and there exists $\widehat{y}$ such that $\widehat{y}^{\mathrm{T}}\widehat{A}_3\widehat{y}<n+1$ and $\widehat{y}_{n+1}^2=1$. Therefore, the condition "$\widehat{A}_3$ is positive semidefinite" is not necessary. In addition, in the statement of [6, Theorem 1], one of the constraints is $y_{n+1}^2=1$. However, in (B.5), the size of the matrix A_i and $\widehat{A_i}$ is $n\times n$ and $(n+1)\times(n+1)$ for $i=1,2,3$, respectively. Therefore, we consider $y\in\mathbb{R}^n$ and $\widehat{y}\in\mathbb{R}^{n+1}$, and change the constraint $y_{n+1}^2=1$ to $\widehat{y}_{n+1}^2=1$.

We now prove that the conditions of Theorem B.1 are satisfied for the constrained Rayleigh quotient optimization problem (B.4).

Lemma B.2. *Suppose* $\|v_0\|<1$, *where* $v_0=(C^{\mathrm{T}})^{\dagger}b$. *Then there exists* $\widehat{y}$ *such that* $\|\widehat{y}\|_N^2=\widehat{y}^{\mathrm{T}}N^{\mathrm{T}}N\widehat{y}<n+1$ *and* $\widehat{y}_{(n-m+1)}=1$.

Proof. Note that $v_0=(C^{\mathrm{T}})^{\dagger}b$ is the minimum norm solution of $C^{\mathrm{T}}v=b$. Let $\widehat{v}=[\sqrt{n}v_0^{\mathrm{T}},1]^{\mathrm{T}}$. Then $\widehat{v}\in\mathcal{N}(\widehat{C})$ and thus there exists $\widehat{y}$ such that $\widehat{v}=N\widehat{y}$ for which we have $\|\widehat{y}\|_N=\|\widehat{v}\|_2<\sqrt{n+1}$, and, at the same time, $\widehat{y}_{(n-m+1)}=\widehat{v}_{(n+1)}=1$. □

By Lemma B.2 and Theorem B.1, the optimization problem (B.4) is equivalent to its dual problem

$$\sup_t \inf_{\widehat{y}^{\mathrm{T}}N^{\mathrm{T}}N\widehat{y}=n+1} \frac{\widehat{y}^{\mathrm{T}}L\widehat{y}+t\widehat{y}_{n-m+1}^2-t}{\widehat{y}^{\mathrm{T}}M\widehat{y}}. \tag{B.8}$$

Since

$$t\widehat{y}_{n-m+1}^2-t=t\widehat{y}_{n-m+1}^2-t\frac{\widehat{y}^{\mathrm{T}}N^{\mathrm{T}}N\widehat{y}}{n+1}=\widehat{y}^{\mathrm{T}}E\widehat{y},$$

(B.8) is equivalent to

$$\sup_t \inf_{\widehat{y}^{\mathrm{T}}N^{\mathrm{T}}N\widehat{y}=n+1} \frac{\widehat{y}^{\mathrm{T}}(L+tE)\widehat{y}}{\widehat{y}^{\mathrm{T}}M\widehat{y}}. \tag{B.9}$$

To transform the dual problem (B.9) to an eigenvalue problem, we first prove that M is positive definite.

Lemma B.3. *Let b be as defined in* (1.1c) *and $b\neq 0$, and N has full column rank, then M is positive definite.*

Proof. It is clear that M is positive semi-definite. We claim that M is nonsingular. Suppose, to the contrary, that M is singular. Then there exists a nonzero x such that $Mx=0$.

We claim that $x_{(n-m+1)}\neq 0$; otherwise suppose $x_{(n-m+1)}=0$ and write $x=\begin{bmatrix}x_1\\0\end{bmatrix}$. It follows from $Mx=0$ that $U^{\mathrm{T}}Ux=0$, implying $x_1=0$ because U has full column rank. Thus $x=0$, a contradiction.

Without loss of generality, we may normalize $x_{(n-m+1)}$ to 1, i.e., $x=\begin{bmatrix} x_1 \\ 1 \end{bmatrix}$. Note that $M=N^{\rm T}N-e_{n-m+1}e_{n-m+1}^{\rm T}$. $Mx=0$ implies $N^{\rm T}Nx=\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. $N^{\rm T}N$ is invertible. We now express $(N^{\rm T}N)^{-1}_{(n-m+1,n-m+1)}$ in two different ways. $N^{\rm T}Nx=\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ yields $x=(N^{\rm T}N)^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and thus

$$1=\begin{bmatrix} 0 \\ 1 \end{bmatrix}^{\rm T}x=\begin{bmatrix} 0 \\ 1 \end{bmatrix}^{\rm T}(N^{\rm T}N)^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix}=(N^{\rm T}N)^{-1}_{(n-m+1,n-m+1)}.$$

On the other hand,

$$N^{\rm T}N=\begin{bmatrix} U^{\rm T}U & U^{\rm T}u \\ u^{\rm T}U & u^{\rm T}u+1 \end{bmatrix}.$$

By the assumption that U has full column rank, $U^{\rm T}U$ is invertible. Then we have

$$\det(N^{\rm T}N)=\det(U^{\rm T}U)\det[(1+u^{\rm T}u-u^{\rm T}U(U^{\rm T}U)^{-1}U^{\rm T}u].$$

According to the relationship between the inverse and the adjoint of a matrix, we find

$$\begin{aligned}(N^{\rm T}N)^{-1}_{(n-m+1,n-m+1)}&=(-1)^{n-m+1+n-m+1}\frac{\det(U^{\rm T}U)}{\det(N^{\rm T}N)}\\&=\frac{\det(U^{\rm T}U)}{\det(U^{\rm T}U)\det[(1+u^{\rm T}u-u^{\rm T}U(U^{\rm T}U)^{-1}U^{\rm T}u]}\\&=\frac{\det(U^{\rm T}U)}{\det(U^{\rm T}U)[1+u^{\rm T}(I-P_U)u]},\end{aligned}$$

where P_U is the orthogonal projection onto $\mathcal{R}(U)$. Therefore, $(N^{\rm T}N)^{-1}_{(n-m+1,n-m+1)}=1$ if and only if $u^{\rm T}(I-P_U)u=0$ implying that u is in the column space of U. Without loss of generality, we may assume the first column of U is u. Now subtract the first column of N from its last column to conclude that e_{n+1} is in the null space of $\widehat{C}$, which contradicts that $b\neq 0$. □

By Lemma B.3 and Courant-Fisher minimax theorem [13, Theorem 8.1.2], finding

$$\inf_{\widehat{y}^{\rm T}N^{\rm T}N\widehat{y}=n+1}\frac{\widehat{y}^{\rm T}(L+tE)\widehat{y}}{\widehat{y}^{\rm T}M\widehat{y}}$$

is equivalent to finding the smallest eigenvalue of $K^{-1}(L+tE)K^{-\rm T}x=\lambda x$, where $M=KK^{\rm T}$ is the Cholesky factorization of M. Therefore, (B.9) is equivalent to

$$\sup_t \lambda_{\min}(L+tE,M). \tag{B.10}$$

Finally, we prove that the maximum value can be obtained, i.e., 'sup' in (B.10) can be replaced by 'max'.

Lemma B.4. *Let $f(t)=\lambda_{\min}(L+tE,M)$. There exits $t_0\in\mathbb{R}$ such that $f(t_0)=\sup_{t\in\mathbb{R}}f(t)$.*

Proof. We prove the claim by showing that

$$\lim_{t\to+\infty}f(t)=\lim_{t\to-\infty}f(t)=-\infty.$$

First, let $v_1\in\mathcal{R}(N)$ with the last component being zero, and set $y_1=N^\dagger v_1$. We have $y_1^{\mathrm{T}}Ey_1=-\frac{\|v_1\|_2^2}{n+1}<0$ and $y_1^{\mathrm{T}}My_1>0$ since M is positive definite. Hence

$$\lim_{t\to+\infty}f(t)=\lim_{t\to+\infty}\inf_{\widehat{y}}\frac{\widehat{y}^{\mathrm{T}}(L+tE)\widehat{y}}{\widehat{y}^{\mathrm{T}}M\widehat{y}}\leqslant\lim_{t\to+\infty}\frac{y_1^{\mathrm{T}}(L+tE)y_1}{y_1^{\mathrm{T}}My_1}\leqslant\lim_{t\to+\infty}t\frac{y_1^{\mathrm{T}}Ey_1}{y_1^{\mathrm{T}}My_1}+\lambda_{\max}(L,M)=-\infty.$$

Recall $v_0=(C^{\mathrm{T}})^\dagger b$ and the assumption that $\|v_0\|<1$. Let $v_2=[\sqrt{n}v_0^{\mathrm{T}},\,1]^{\mathrm{T}}$. Clearly $v_2\in\mathcal{R}(N)$ and let $y_2=N^\dagger v_2$. We have $y_2^{\mathrm{T}}Ey_2=-\frac{n\|v_0\|_2^2}{n+1}+1-\frac{1}{n+1}>0$ since $\|v_0\|<1$ and $y_2^{\mathrm{T}}My_2>0$ since M is positive definite. Hence

$$\lim_{t\to-\infty}f(t)=\lim_{t\to-\infty}\inf_{\widehat{y}}\frac{\widehat{y}^{\mathrm{T}}(L+tE)\widehat{y}}{\widehat{y}^{\mathrm{T}}M\widehat{y}}\leqslant\lim_{t\to-\infty}\frac{y_2^{\mathrm{T}}(L+tE)y_2}{y_2^{\mathrm{T}}My_2}\leqslant\lim_{t\to-\infty}t\frac{y_2^{\mathrm{T}}Ey_2}{y_2^{\mathrm{T}}My_2}+\lambda_{\max}(L,M)=-\infty.$$

Therefore, there exits $t_1<0$ such that $f(t)<f(0)$ for $t<t_1$ and there exits $t_2>0$ such that $f(t)<f(0)$ for when $t>t_2$. Therefore

$$\sup_{t\in\mathbb{R}}f(t)=\sup_{t\in[t_1,t_2]}f(t).$$

Because $f(t)=\lambda_{\min}(L+tE,M)$ is a continuous function [38], there exists $t_0\in[t_1,t_2]$ such that $f(t_0)=\sup_{t\in\mathbb{R}}f(t)$. □

In conclusion, we have shown that CRQopt (1.1) is equivalent to the eigenvalue optimization problem (B.2).

C CRQPACK

The Lanczos algorithm for solving CRQopt (1.1) described in this paper has been implemented in MATLAB. In the spirit of reproducible research, MATLAB scripts of the implementation of the Lanczos algorithm and the data that used to generate numerical results presented in this paper are available in CRQPACK at `https://github.com/yunshenzhou/CRQPACK.git`. CRQPACK consists of three folders:

1. `src`: the source code for solving CRQopt (1.1). It consists of four functions `CRQ_Lanczos`, `QEPmin`, `LGopt` and `rLGopt`. `CRQ_Lanczos` is the driver and calls `QEPmin` and `LGopt`. `LGopt` is dependent on `rLGopt`. In addition, we also provide two other drivers for solving CRQopt (1.1), namely `CRQ_explicit` for the direct method [10] and `CRQ_ppm` for the projected power method [41].

2. `synthetic`: the drivers for numerical examples in Section 3.6. `correct.m` and `QEPres.m` are for the examples in Sections 3.2 and 3.5, respectively. `CRQsharp.m` is used to generate the plots for Example 3.3 on error bounds in (3.49a) and (3.49b), while `CRQnotsharp.m` on the error bounds (3.49a) and (3.49b).

3. `imagecut`: the code for constrained image segmentation. It has three subfolders: `examples` contains the drivers, `data` contains image data including prior labeling information, and `auxiliary` contains program to generate the matrices A, C, and vector b of CRQopt (1.1).

References

[1] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[2] J. R. Bunch, Ch. P. Nielsen, and D. C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numer. Math.*, 31:31–48, 1978.

[3] S. E. Chew and N. D. Cahill. Semi-supervised normalized cuts for image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1716–1723, 2015.

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

[5] N. R. Draper. "Ridge analysis" of response surfaces. *Technometrics*, 5(4):469–479, 1963.

[6] A. Eriksson, C. Olsson, and F. Kahl. Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. *Journal of Mathematical Imaging and Vision*, 39(1):45–61, 2011.

[7] D. Fong and M. Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM J. Sci. Comput.*, 33(5):2950–2971, 2011.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements Of Statistical Learning*. Springer, 2001.

[9] W. Gander. Least squares with a quadratic constraint. *Numer. Math.*, 36:291–307, 1981.

[10] W. Gander, G. H. Golub, and U. von Matt. A constrained eigenvalue problem. *Linear Algebra Appl.*, 114-115:815–839, 1989.

[11] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Academic Press, New York, 1982.

[12] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[13] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, 4th edition, 2013.

[14] Gene H. Golub, Zhenyue Zhang, and Hongyuan Zha. Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner-outer iterations. *Linear Algebra Appl.*, 309(1):289–306, 2000.

[15] G. Golubr. Some modified matrix eigenvalue problems. *SIAM Rev.*, 15:318–334, 1973.

[16] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM J. Optim.*, 9(2):504–525, 1999.

[17] W. W. Hager. Minimizing a quadratic over a sphere. *SIAM J. Optim.*, 12(1):188–208, 2001.

[18] C. Jiang, H. Xie, and Z. Bai. Robust and efficient computation of eigenvectors in a generalized spectral method for constrained clustering. In *Artificial Intelligence and Statistics*, pages 757–766, 2017.

[19] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 290–297, 2003.

[20] J. Lampe and H. Voss. On a quadratic eigenproblem occurring in regularized total least squares. *Computational Statistics & Data Analysis*, 52(2):1090–1102, 2007.

[21] R.-C. Li. Solving secular equations stably and efficiently. Technical Report UCB//CSD-94-851, Computer Science Division, Department of EECS, University of California at Berkeley, 1993.

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

[23] R.-C. Li. On Meinardus' examples for the conjugate gradient method. *Math. Comp.*, 77(261):335–352, 2008.

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

[25] W. Liu, G. Hua, and J. R. Smith. Unsupervised one-class learning for automatic outlier removal. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3826–3833, June 2014.

[26] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67, 2007.

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

[28] M. E. J. Newman. Spectral methods for community detection and graph partitioning. *Phys. Rev. E*, 88:042822, 2013.

[29] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71, 1982.

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

[31] D. Paurat, D. Oglic, and T. Gärtner. Supervised PCA for interactive data analysis. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS) 2nd Workshop on Spectral Learning*, 2013.

[32] W. Pentney and M. Meila. Spectral clustering of biological sequence data. In *Association for the Advancement of Artificial Intelligence*, pages 845–850, 2005.

[33] F. Rendl and H. Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Math. Program., Ser. A*, 77(1):273–299, 1997.

[34] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, Manchester, UK, 1992.

[35] M. A. Saunders. Solution of sparse rectangular systems using LSQR and CRAIG. *BIT Numer. Math.*, 35(4):588–604, 1995.

[36] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[37] D. M. Sima, S. Van Huffel, and G. H. Golub. Regularized total least squares based on quadratic eigenvalue problem solvers. *BIT Numerical Mathematics*, 44(4):793–812, 2004.

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

[39] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[40] X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.

[41] L. Xu, W. Li, and D. Schuurmans. Fast normalized cut with linear constraints. In *2009 IEEE*

Conference on Computer Vision and Pattern Recognition, pages 2866–2873, June 2009.

[42] S. X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.

[43] L.-H. Zhang, C. Shen, and R.-C. Li. On the generalized Lanczos trust-region method. *SIAM J. Optim.*, 27(3):2110–2142, 2017.