Journal of Scientific Computing (2024) 100:89
https://doi.org/10.1007/s10915-024-02636-9

Spectrally Constrained Optimization

Casey Garner[1] · **Gilad Lerman**[1] · **Shuzhong Zhang**[2]

Received: 20 December 2023 / Revised: 15 July 2024 / Accepted: 19 July 2024 /
Published online: 6 August 2024

Abstract
We investigate how to solve smooth matrix optimization problems with general linear inequality constraints on the eigenvalues of a symmetric matrix. We present solution methods to obtain exact global minima for linear objective functions, i.e., $F(\boldsymbol{X}) = \langle \boldsymbol{C}, \boldsymbol{X} \rangle$, and perform exact projections onto the eigenvalue constraint set. Two first-order algorithms are developed to obtain first-order stationary points for general non-convex objective functions. Both methods are proven to converge sublinearly when the constraint set is convex. Numerical experiments demonstrate the applicability of both the model and the methods.

Keywords Eigenvalue optimization · Constrained optimization · Frank-Wolfe algorithm · Non-smooth analysis · Matrix completion

Mathematics Subject Classification 90C26 · 90C52 · 65K10 · 68W40

1 Introduction

Constrained matrix optimization is an essential aspect of machine learning [14], matrix factorization and completion [4, 7, 32], semidefinite programming [53], robust subspace recovery [33, 54] and covariance estimation [15]. There are two types of constraints to consider in matrix optimization: coordinate constraints and spectral constraints. Coordinate constraints impose prohibitions on the entries of the matrix, such as the diagonal elements must equal one or the row and column sums must be unity. Spectral constraints force the eigenvalues or singular values to satisfy certain conditions. For example, assuming a matrix is positive semidefinite enforces non-negativity on the eigenvalues. The literature is replete with a diverse array of coordinate constraints, but the variety of spectral constraints is limited. Given the power of constrained matrix optimization, the lack of general spectral constraints, with corresponding theory and algorithmic development, is a knowledge gap worth filling.

The goal of this paper is to explore spectrally constrained matrix optimization (SCO) and develop theory and algorithms to solve new models with general spectral constraints. We begin with the following spectrally constrained eigenvalue model

✉ Casey Garner
garne214@umn.edu

[1] School of Mathematics, University of Minnesota, Minneapolis, MN, USA

[2] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, USA

$$\begin{aligned} \min\ & F(\boldsymbol{X}) \\ \text{s.t. } & \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b} \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n} \end{aligned} \tag{SCO-Eig}$$

where $\mathcal{S}^{n\times n}$ denotes the set of real n-by-n symmetric matrices, $F : \mathcal{S}^{n\times n} \to \mathbb{R}$ is continuously differentiable, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{\lambda}(\boldsymbol{X})$ is the vector of eigenvalues of $\boldsymbol{X}$ in descending order, i.e., $\lambda_1(\boldsymbol{X}) \geq \lambda_2(\boldsymbol{X}) \geq \cdots \geq \lambda_n(\boldsymbol{X})$. (SCO-Eig) allows general linear inequality constraints on the eigenvalues; therefore, the model encompasses traditional spectral constraints, such as non-negativity, while significantly boosting the modeling power of the practitioner. As an example, low-rank conditions are important and popular in matrix optimization [39, 62]. The general low-rank model, $\min\{F(\boldsymbol{X}) \mid \operatorname{rank}(\boldsymbol{X}) \leq k,\ \boldsymbol{X} \in \mathcal{S}_+^{n\times n}\}$ where $\mathcal{S}_+^{n\times n}$ denotes the set of real n-by-n positive semidefinite matrices, can be written in the form of (SCO-Eig) as

$$\min\left\{F(\boldsymbol{X}) \mid \lambda_n(\boldsymbol{X}) \geq 0, \lambda_{k+1}(\boldsymbol{X}) \leq 0,\ \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\}$$

or approximated as $\min\left\{F(\boldsymbol{X}) \mid \lambda_i(\boldsymbol{X}) \in [0,\delta], i = k+1,\ldots,n,\ \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\}$, where $\delta \geq 0$ controls the accuracy of the approximation.

This paper presents the first study of (SCO-Eig) and is a departure from the current corpus devoted to matrix optimization. Our approach offers a change of perspective which has somehow escaped scrutiny by the community. To demarcate the novelty of our framework, we embark on a brief excursion into the literature.

1.1 Literature Review

Matrix optimization with an emphasis on the spectrum has been a subject of rigorous exploration for decades. These efforts have amassed a sizable body of work known as *eigenvalue optimization* [11, 36, 37, 42, 45–47, 51, 59]. The main focus of these papers is minimizing functions of the eigenvalues or singular values of a constrained parameterized matrix. A general framework often seen in eigenvalue optimization models is

$$\begin{aligned} \min\ & f(\boldsymbol{\lambda}(\mathcal{A}(\boldsymbol{x}))) \\ \text{s.t. } & \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^m, \end{aligned} \tag{1}$$

where $\mathcal{A} : \mathbb{R}^m \to \mathcal{S}^{n\times n}$ is a smooth map and Ω could be a strict subset of $\mathbb{R}^m$. Many of these works focus on a specific form of the objective function in (1), such as minimizing the maximum eigenvalue of a parameterized matrix [45] or minimizing a weighted-sum of the k-largest eigenvalues or singular values [11, 47, 59]. According to Overton [47], the work of Cullum, Donath and Wolfe in 1975 seems to be the first study of minimizing the sum of the k-largest eigenvalues of a parameterized matrix. In particular, Cullum et al. investigated minimizing $f_\kappa(\boldsymbol{x}) = \sum_{i=1}^{\kappa} \lambda_i(\mathcal{A}(\boldsymbol{x}))$ where $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{A}_0 + \operatorname{diag}(\boldsymbol{x})$ for a fixed matrix $\boldsymbol{A}_0$. In the case of Hermitian matrices, Mengi et al. [42] provide a general framework similar to (1), and Kangal et al. [29] consider an infinite dimensional problem by minimizing the k-th largest eigenvalue of a compact self-adjoint operator. Early overviews of eigenvalue optimization are provided by Lewis and Overton [36, 37].

The seminal work of Cullum, Donath and Wolfe focused on understanding the non-smoothness of f_κ, and the study of non-smooth functions composed of eigenvalues and singular values has remained a common direction of inquiry. We see this in the works of Overton [45–47] and Lewis [23, 34, 35, 38] where much discussion surrounds computing the subdifferentials of functions whose arguments are eigenvalues and singular values. A

particularly important setting where subdifferentials and derivatives are readily available in eigenvalue optimization is that of *spectral functions*. Spectral functions are a staple in the eigenvalue optimization literature [1, 34, 38]. The function $F : \mathcal{S}^{n\times n} \to \mathbb{R}$ is a spectral function (over $\mathcal{S}^{n\times n}$) if $F(\boldsymbol{X}) = f(\boldsymbol{\lambda}(\boldsymbol{X}))$ for some *symmetric function* $f : \mathbb{R}^n \to \mathbb{R}$ where f *symmetric* means $f(\boldsymbol{x}) = f(\boldsymbol{P}\boldsymbol{x})$ for all permutation matrices $\boldsymbol{P} \in \mathbb{R}^{n\times n}$. The key feature of spectral functions is they are independent of the order of the eigenvalues, e.g., $F(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}) = \sum_{i=1}^{n} \lambda_i(\boldsymbol{X})$ is a spectral function. Example 7.13 in [1] presents a number of common spectral functions encountered in the literature.

Studies on how to avoid non-smoothness also exist. Shapiro and Fan [51] detailed how to avoid some issues of non-smoothness at repeated eigenvalues by proving under certain conditions the set $\{\boldsymbol{x} \in \mathbb{R}^m \mid \lambda_1(\mathcal{A}(\boldsymbol{x})) = \cdots = \lambda_k(\mathcal{A}(\boldsymbol{x}))\}$ is a smooth manifold near a point $\boldsymbol{x}^*$ where $\lambda_1(\mathcal{A}(\boldsymbol{x}^*))$ has multiplicity and $\mathcal{A}$ is a smooth map from $\mathbb{R}^m$ to $\mathcal{S}^{n\times n}$.

1.2 Contributions: A New Perspective

A common thread present in the eigenvalue optimization literature is minimizing objective functions solely dependent on the spectrum of the decision matrix. As a result, these papers fixate on exploring the non-smoothness of these objectives. This has led to beautiful mathematics and robust application; however, the success of these models has simultaneously limited the scope of the investigation. Constrained matrix optimization revolves around coordinate and spectral constraints, and the current eigenvalue optimization literature has not expanded our understanding of general spectral constraints.

Our work on (SCO-Eig) begins a novel study to understand eigenvalues being present functionally in the constraint set. Ours is a perspective shift, instead of focusing on eigenvalues in the objective, we focus on eigenvalues in the constraint, and, to our surprise, this perspective seems uncommon. Almost no papers consider general functional constraints on the eigenvalues, e.g., $g(\lambda(\boldsymbol{X})) \leq 0$, and those that do restrict themselves to spectral functions. In the very recent work [39], which appeared while we were finishing this paper, the authors allow spectral functions in the constraint set they consider; however, they do not go beyond this paradigm and they further require convexity of the spectral functions. We do not assume the constraint set in (SCO-Eig) is composed of spectral functions, and, unlike [39], we develop approaches to solve our model without convexification. Another work which considers general functional eigenvalue constraints is [28]. In this paper, the authors investigate error bound conditions for constraint sets of the form $\left\{\boldsymbol{X} \in \mathcal{S}^{n\times n} \mid f(\boldsymbol{\lambda}(\boldsymbol{X})) \leq 0\right\}$ where $f : \mathbb{R}^n \to \mathbb{R}$, but they did not utilize their results to investigate matrix optimization. A paper which appeared following the completion of this work by Ito and Lourenço [26] is the closest to our study. They approached the same type of problem we consider using a different mathematical framework. Their work supports our belief this is the genesis of a fruitful optimization model.

It is important to note a comprehensive study of (SCO-Eig) must depart from the current literature around spectral functions. First, the constraint set we consider cannot be obtained via spectral functions except in special circumstances. Second, the objective function we consider is never assumed to be a spectral function. For example, even in the simple case of a linear objective function, which we study in Sect. 4, the objective function is not, except in special instances, a spectral function; therefore, one cannot simply utilize the current literature to directly answer the questions we are posing, so we sought to expand the current state-of-affairs to address new challenges.

The main contribution of this work is the first study of a matrix optimization model with general linear inequality constraints on the eigenvalues. We prove in the case of a linear objective function that a global minimum of (SCO-Eig) can be computed regardless of the non-convexity of the constraint set, and we prove how to perform optimal projections onto the constraint set. With these results, we develop two first-order algorithms to compute stationary points to (SCO-Eig) and provide a proof of their sublinear convergence rate. Numerical experiments demonstrate the applicability of our procedures.

1.3 Organization

The organization of the paper is as follows. Section 2 presents and proves some facts about the eigenvalue constraint set, such as its connectedness. Section 3 provides a discussion of necessary optimality conditions for (SCO-Eig) and an approximated version of the model which avoids the non-smoothness associated with repeated eigenvalues. Sections 4 and 5 demonstrate we can solve essential instantiations of (SCO-Eig) which form crucial steps in classical optimization algorithms. Namely, Sect. 4 proves solving (SCO-Eig) with a linear objective function can be done exactly, regardless of the non-convexity of the constraint, and only requires solving a linear program and performing a spectral decomposition. Section 5 solves the projection problem onto the eigenvalue constraint set. From these results, in Sect. 6 we present a projected gradient method and a Frank-Wolfe algorithm which obtain first-order ϵ-stationary points to (SCO-Eig) when the constraint set is convex and the objective function is smooth though possibly non-convex. Section 7 presents the convergence analysis for our presented algorithms, and Sect. 8 displays the results of numerical experimentation; Sect. 8.1 applies both methods to the question of determining a preconditioning matrix for solving linear systems; Sect. 8.2 investigates the solving of systems of quadratic equations with our projected gradient method while Sect. 8.3 presents an example in matrix completion. We conclude the paper in Sect. 9 with directions for future inquiries.

In this paper, scalars, vectors and matrices are denoted by lower case, bold lower case, and bold upper case letters respectively, e.g., x, $\boldsymbol{y}$ and $\boldsymbol{Z}$. Given the positive integer k, $[k] := \{1, \ldots, k\}$. If $\boldsymbol{x} \in \mathbb{R}^n$, $\mathrm{Diag}(\boldsymbol{x})$ denotes the diagonal matrix whose diagonal is $\boldsymbol{x}$. The set of real n-by-n symmetric, positive semidefinite, and positive definite matrices are denoted $\mathcal{S}^{n\times n}$, $\mathcal{S}^{n\times n}_{+}$, and $\mathcal{S}^{n\times n}_{++}$ respectively, and $\mathcal{O}(n) := \left\{\boldsymbol{X} \in \mathbb{R}^{n\times n} \mid \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}\right\}$ denotes the set of orthogonal matrices where $\boldsymbol{I}$ is the identity matrix.

2 The Eigenvalue Constraint Set

The paradigm shift we are pursuing rests on the eigenvalues being functionally present in the constraints. Due to the limited attention received by such settings, we begin by proving some general facts about the constraint set in (SCO-Eig), which we denote as

$$\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) := \left\{\boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b}\right\}. \tag{2}$$

We shall desire to take advantage of writing the constraint without ordering the eigenvalues. To this end, let $\boldsymbol{\lambda}_{\boldsymbol{X}} \in \mathbb{R}^n$ be the unordered vector of eigenvalues for $\boldsymbol{X} \in \mathcal{S}^{n\times n}$; we define the matrix $\boldsymbol{D}_n \in \mathbb{R}^{(n-1)\times n}$ to enforce the descending order of the eigenvalues, i.e., the i-th row of $\boldsymbol{D}_n$ is $(\boldsymbol{e}_{i+1} - \boldsymbol{e}_i)^\top$ where $\boldsymbol{e}_i \in \mathbb{R}^n$ has a one in the i-th place and zeros elsewhere. Thus, an equivalent manner of writing (2) is $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) = \left\{\boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \boldsymbol{A}\boldsymbol{\lambda}_{\boldsymbol{X}} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda}_{\boldsymbol{X}} \leq \boldsymbol{0}\right\}$.

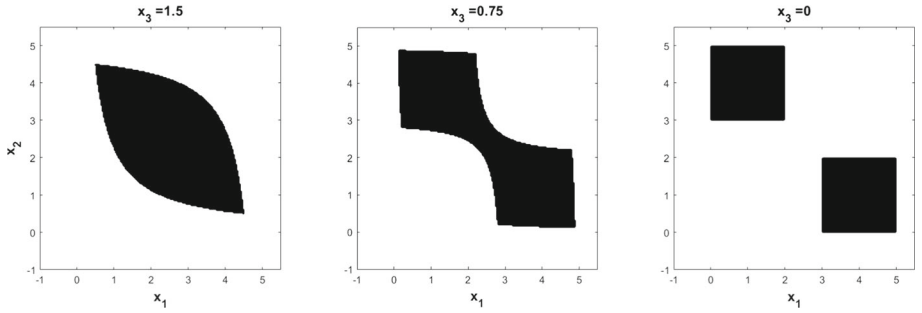

Fig. 1 Demonstration of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ in Example 1. Each subplot shows the black regions of coordinates x_1 and x_2 of $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ for a fixed value of x_3, which is specified above each subplot

Our first fact demonstrates the unsurprising but key truth that $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ has the potential to be convex and non-convex as a function of the problem's data.

Fact 1 *$\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ can be both convex and non-convex dependent on $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.*

Proof If $\boldsymbol{A} = -\boldsymbol{I}$ and $\boldsymbol{b} = \boldsymbol{0}$, then $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) = \mathcal{S}_+^{n \times n}$ which is convex. Let $\boldsymbol{A} \in \mathbb{R}^{2 \times 2}$ and $\boldsymbol{b} \in \mathbb{R}^2$ enforce the constraint $\boldsymbol{X} \in \mathcal{S}^{2 \times 2}$ with $\lambda_1(\boldsymbol{X}) \geq 3$ and $\lambda_2(\boldsymbol{X}) \leq 1$. One can easily produce examples where convex combinations of matrices satisfying these constraints fail these conditions. For example, the matrices

$$\boldsymbol{X}_1 = \begin{pmatrix} 35 & 15 \\ 15 & 6 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X}_2 = \begin{pmatrix} 4 & 17 \\ 17 & 63 \end{pmatrix}$$

satisfy the eigenvalue constraints; however, $\boldsymbol{Y} := \frac{1}{2}(\boldsymbol{X}_1 + \boldsymbol{X}_2)$ fails the condition $\lambda_2(\boldsymbol{Y}) \leq 1$. □

We further explore the non-convexity of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ with an example.

Example 1 Let $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) = \left\{\boldsymbol{X} \in \mathcal{S}^{2 \times 2} \mid \boldsymbol{\lambda}_1(\boldsymbol{X}) \in [3, 5],\ \boldsymbol{\lambda}_2(\boldsymbol{X}) \in [0, 2]\right\}$. This set is non-convex. To visualize this, we take slices of the constraint set. Let

$$\boldsymbol{X} = \begin{pmatrix} x_1 & x_3 \\ x_3 & x_2 \end{pmatrix} \quad \text{with} \quad x_1, x_2, x_3 \in \mathbb{R}.$$

Fixing x_3 for different values, we plot in Fig. 1 the values of x_1 and x_2 such that $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$. The three subplots in Fig. 1 clearly display the non-convexity of the set. When $x_3 = 0$, the disjoint interval constraints become evident in the two square regions of the third subplot.

Though the constraint set in (SCO-Eig) may be non-convex, the feasible region, however, is always connected.

Fact 2 *$\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ is a connected subset of $\mathcal{S}^{n \times n}$.*

Proof Let $\boldsymbol{X}_1 = \boldsymbol{Q}_1 \boldsymbol{\Lambda}_1 \boldsymbol{Q}_1^\top$ and $\boldsymbol{X}_2 = \boldsymbol{Q}_2 \boldsymbol{\Lambda}_2 \boldsymbol{Q}_2^\top$ be arbitrary elements of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$. Without loss of generality, we may assume $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in SO(n) := \{X \in \mathcal{O}(n) \mid \det(\boldsymbol{X}) = 1\}$. Moreover, we may assume without loss of generality the diagonal elements of $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are in descending order. Using the fact $SO(n)$ is path connected, there exists a continuous map $\boldsymbol{G} : [0, 1] \to SO(n)$ such that $\boldsymbol{G}(0) = \boldsymbol{Q}_1$ and $\boldsymbol{G}(1) = \boldsymbol{Q}_2$. Thus, we can define the continuous parameterization $\boldsymbol{X} : [0, 1] \to \mathcal{S}^{n \times n}$ such that

$$\boldsymbol{X}(t) = \boldsymbol{G}(t)\left(\boldsymbol{\Lambda}_2 + (1 - t)\left(\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\right)\right)\boldsymbol{G}(t)^\top.$$

By construction, $\boldsymbol{X}(0) = \boldsymbol{X}_1$ and $\boldsymbol{X}(1) = \boldsymbol{X}_2$, and the convexity of

$$\left\{\boldsymbol{\lambda} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0}\right\}$$

ensures $\boldsymbol{X}(t) \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ for all $t \in [0, 1]$. □

Remark 1 Figure 1, especially the third subplot in Fig. 1, might seem to suggest a counter-example refuting the connectedness of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$. However, this is not the case. The stills in Fig. 1 represent slices of the feasible region and paths connecting points in $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ are not restricted to slices.

The capacity for non-convexity in the constraint is the trade-off incurred by removing the eigenvalues from the objective. Potential non-smoothness of the objective function associated with the eigenvalues has been replaced with potential non-convexity associated with the eigenvalues in the constraint. Constrained problems over non-convex sets are exceptionally challenging; therefore, the cases where $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ is convex are of interest. The next theorem provides a verifiable condition which ensures the convexity of the constraint and proves the convexity of common eigenvalue constraints such as $\mathcal{S}_+^{n\times n}$ and condition number constraints [52, 58]. To prove the result we require the following proposition

Proposition 1 *If $\alpha_1 \geq \ldots \geq \alpha_n$, then $f(\boldsymbol{X}) = \sum_{i=1}^n \alpha_i \lambda_i(\boldsymbol{X})$ is convex.*

Proof It is well-known $g_k(\boldsymbol{X}) := \sum_{i=1}^k \lambda_i(\boldsymbol{X})$ and $h_p(\boldsymbol{X}) := \sum_{j=0}^p \lambda_{n-j}(\boldsymbol{X})$ are convex and concave functions respectively for any $1 \leq k \leq n$ and $0 \leq p \leq n-1$ [3]. From this and simple operations which preserve convexity it follows $g_\alpha(\boldsymbol{X}) := \sum_{i=1}^n \alpha_i \lambda_i(\boldsymbol{X})$ and $h_\beta(\boldsymbol{X}) := \sum_{j=0}^{n-1} \beta_{n-j} \lambda_{n-j}(\boldsymbol{X})$ with $\alpha_1 \geq \ldots \geq \alpha_n \geq 0$ and $\beta_n \leq \ldots \leq \beta_1 \leq 0$ are convex. Thus,

$$g_\alpha(\boldsymbol{X}) + h_\beta(\boldsymbol{X}) = \sum_{i=1}^n (\alpha_i + \beta_i)\lambda_i(\boldsymbol{X})$$

is convex. The equivalence of the sets $\{\boldsymbol{x} \in \mathbb{R}^n \mid x_1 \geq \ldots \geq x_n\}$ and

$$\left\{\boldsymbol{y} + \boldsymbol{z} \in \mathbb{R}^n \mid y_1 \geq \ldots \geq y_n \geq 0,\ z_n \leq \ldots \leq z_1 \leq 0\right\}$$

completes the argument. □

Utilizing Proposition 1, we present a general condition guaranteeing the convexity of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$.

Theorem 1 *If each row of $\boldsymbol{A}$ is an element of $\mathbb{R}^n_{\geq} := \{\boldsymbol{x} \in \mathbb{R}^n \mid x_1 \geq \ldots \geq x_n\}$, $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ is convex.*

Proof Assume each row of $\boldsymbol{A}$ is an element of $\mathbb{R}^n_{\geq}$. For $i = 1, \ldots, m$ we define $f_i(\boldsymbol{X}) := \sum_{j=1}^n A_{ij}\lambda_j(\boldsymbol{X})$. Since each f_i is convex by Proposition 1, the level sets of each f_i are convex. The result then follows because $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ is the intersection of the level sets of the f_i's. □

3 General Optimality Conditions

The non-smoothness of the eigenvalues necessitates a discussion of concepts from non-smooth analysis. Consider an open set Ω of $\mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$ locally Lipschitz on Ω, i.e.,

given $\boldsymbol{x} \in \Omega$ there exists $L_{\boldsymbol{x}}$ and $\delta_{\boldsymbol{x}} > 0$ such that if $\|\boldsymbol{y} - \boldsymbol{x}\| \leq \delta_{\boldsymbol{x}}$ then $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L_{\boldsymbol{x}}\|\boldsymbol{x} - \boldsymbol{y}\|$. The (*radial*) *directional derivative* of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{d} \in \mathbb{R}^n$ is

$$f'(\boldsymbol{x}; \boldsymbol{d}) := \lim_{t \to 0_+} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t},$$

when the limit exists. A relaxed version of the directional derivative is the *Clarke directional derivative*. The *Clarke directional derivative* of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{d}$ is

$$f^C(\boldsymbol{x}; \boldsymbol{d}) := \limsup_{(t, \boldsymbol{y}) \to (0_+, \boldsymbol{x})} \frac{f(\boldsymbol{y} + t\boldsymbol{d}) - f(\boldsymbol{y})}{t}.$$

Unlike $\boldsymbol{d} \mapsto f'(\boldsymbol{x}; \boldsymbol{d})$, $\boldsymbol{d} \mapsto f^C(\boldsymbol{x}; \boldsymbol{d})$ is guaranteed to be a finite convex function for all $\boldsymbol{x} \in \Omega$ (see Sect. 5.1 in [48]). The *Clarke subdifferential* of f at $\boldsymbol{x}$ is

$$\partial_C f(\boldsymbol{x}) := \left\{ \boldsymbol{s} \in \mathbb{R}^n \mid \langle \boldsymbol{s}, \boldsymbol{d} \rangle \leq f^C(\boldsymbol{x}; \boldsymbol{d}),\ \forall \boldsymbol{d} \in \mathbb{R}^n \right\}.$$

The Clarke subdifferential is a non-empty, compact and convex set, and these properties make it a frequently utilized generalization of differentiation for non-smooth functions. The definitions presented here for Clarke subdifferentials can be found in Sect. 2 of [23]. It is often necessary to assume a function is *regular* at a point for certain results to hold. Different definitions of regularity exist, but in this section we say f is *regular* at $\boldsymbol{x} \in \Omega$ if

$$f^C(\boldsymbol{x}; \boldsymbol{d}) = \liminf_{(t, \boldsymbol{v}) \to (0_+, \boldsymbol{d})} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}$$

for all $\boldsymbol{d} \in \mathbb{R}^n$ (see Definition 5.46 and Proposition 4.3 in [48]). This notion of regularity focuses on the equality of two slightly different generalizations of the directional derivative of f. One deals with perturbing the point $\boldsymbol{x}$ while the other perturbs the direction $\boldsymbol{d}$. Convex functions are regular at all points in their domain and if f is continuously differentiable at $\boldsymbol{x}$ then it is also regular at $\boldsymbol{x}$. A thorough discourse on Clarke subdifferentials and regularity is presented in [48]. The interested reader should consult this text and the references therein for further details. We now present general necessary conditions for optimal solutions to (SCO-Eig).

Theorem 2 *Assume $F : \mathcal{S}^{n \times n} \to \mathbb{R}$ is continuously differentiable and $g_i(\boldsymbol{X}) := \boldsymbol{a}_i^\top \boldsymbol{\lambda}(\boldsymbol{X}) - b_i$ for $i \in [p]$. Let $\boldsymbol{X}^*$ be a local minimizer of*

$$\min \left\{ F(\boldsymbol{X}) \mid g_i(\boldsymbol{X}) \leq 0,\ i \in [p],\ \boldsymbol{X} \in \mathcal{S}^{n \times n} \right\}$$

and $\mathcal{I}(\boldsymbol{X}^*) := \left\{ i \in [p] : g_i(\boldsymbol{X}^*) = 0 \right\}$. Suppose

$$\boldsymbol{t} \in \mathbb{R}^p_+,\ t_i = 0\ \forall i \in [p] \setminus \mathcal{I}(\boldsymbol{X}^*),\ 0 \in t_1 \partial_C g_1(\boldsymbol{X}^*) + \cdots + t_p \partial_C g_p(\boldsymbol{X}^*) \implies \boldsymbol{t} = \boldsymbol{0}$$

and each g_i is regular at $\boldsymbol{X}^*$, then there exist multipliers $\mu \in \mathbb{R}^p_+$ such that, $\mu_i g_i(\boldsymbol{X}^*) = 0$ for all i and,

$$\begin{aligned} 0 \in \nabla F(\boldsymbol{X}^*) &+ \mu_1 \sum_{i=1}^n a_{1i}\, conv \left\{ \boldsymbol{v}\boldsymbol{v}^\top \mid \boldsymbol{v} \in \mathbb{E}_i(\boldsymbol{X}^*),\ \|\boldsymbol{v}\| = 1 \right\} \\ &+ \cdots + \mu_p \sum_{i=1}^n a_{pi}\, conv \left\{ \boldsymbol{v}\boldsymbol{v}^\top \mid \boldsymbol{v} \in \mathbb{E}_i(\boldsymbol{X}^*),\ \|\boldsymbol{v}\| = 1 \right\} \end{aligned} \tag{3}$$

where $\mathbb{E}_i(\boldsymbol{X}^*)$ denotes the eigenspace of $\boldsymbol{X}^*$ corresponding to the i-th largest eigenvalue of $\boldsymbol{X}^*$ and $conv(S)$ is the convex hull of a given set S.

Proof Using the framework of Corollary 5.54 in [48], let $f = F$ and g_i be as defined. Since $\boldsymbol{\lambda}(\cdot)$ is globally Lipschitz continuous, g_i is locally Lipschitz for all $\boldsymbol{X} \in \mathcal{S}^{n\times n}$. By the assumption F is continuously differentiable, it follows $\partial_C F(\boldsymbol{X}) = \{\nabla F(\boldsymbol{X})\}$ for all $\boldsymbol{X}$. The proof is completed by computing the Clarke subdifferential of the g_i's. Since g_i is regular at $\boldsymbol{X}$,

$$\begin{aligned}\partial_C g_i(\boldsymbol{X}) &= \sum_{j=1}^{n} \partial_C \left(a_{ij}\boldsymbol{\lambda}_j(\cdot)\right)(\boldsymbol{X}) \\ &= \sum_{j=1}^{n} a_{ij}\partial_C \boldsymbol{\lambda}_j(\boldsymbol{X}) \\ &= \sum_{j=1}^{n} a_{ij}\mathrm{conv}\left\{\boldsymbol{v}\boldsymbol{v}^\top \mid \boldsymbol{v} \in \mathbb{E}_j(\boldsymbol{X}),\ \|\boldsymbol{v}\| = 1\right\} \end{aligned} \tag{4}$$

where the first and second equalities follow by Theorem 5.51 and Proposition 5.9 in [48] respectively. The last equality is due to Theorem 5.3 in [23]. □

Theorem 2 provides general optimality conditions for (SCO-Eig). Note, if $g_i(\boldsymbol{X}) = \boldsymbol{a}_i^\top \lambda(\boldsymbol{X}) - b_i$ for $i \in [m]$ where $\boldsymbol{a}_i$ is the i-th row of $\boldsymbol{A}$, then $\mathcal{S}_\lambda(\boldsymbol{A}, \boldsymbol{b}) = \left\{\boldsymbol{X} \in \mathcal{S}^{n\times n} \mid g_i(\boldsymbol{X}) \le 0,\ i \in [m]\right\}$. The assumptions required for Theorem 2 to hold are substantial including: a constraint qualification, regularity and non-trivial convex hulls of eigenspaces; however, if the local minimizer has unique eigenvalues, the necessary conditions simplify immensely.

Theorem 3 *Assume $F : \mathcal{S}^{n\times n} \to \mathbb{R}$ is continuously differentiable. Define $g_i(\boldsymbol{X}) := \boldsymbol{a}_i^\top \boldsymbol{\lambda}(\boldsymbol{X}) - b_i$ for $i \in [p]$. Let $\boldsymbol{X}^*$ be a local minimizer of*

$$\min_{\boldsymbol{X}\in\mathcal{S}^{n\times n}} \{F(\boldsymbol{X}) \mid g_i(\boldsymbol{X}) \le 0,\ i \in [p]\}$$

with no repeated eigenvalues. If $\{\boldsymbol{a}_i \mid i \in \mathcal{I}(\boldsymbol{X}^*)\}$ is a linearly independent set, then there exist multipliers $\mu \in \mathbb{R}^p_+$ such that,

$$\nabla F(\boldsymbol{X}^*) + \sum_{i=1}^{p} \mu_i^* \boldsymbol{V}^* \mathrm{Diag}(\boldsymbol{a}_i)(\boldsymbol{V}^*)^\top = \boldsymbol{0},$$

and $\mu_i^* g_i(\boldsymbol{X}^*) = 0$ for all i where $\boldsymbol{X}^* = \boldsymbol{V}^*\mathrm{Diag}(\boldsymbol{\lambda}(\boldsymbol{X}^*))(\boldsymbol{V}^*)^\top$.

Proof Since $\boldsymbol{X}^*$ has no repeated eigenvalues, g_i is regular at $\boldsymbol{X}^*$. The uniqueness of the eigenvalues of $\boldsymbol{X}^*$ imply there is a neighborhood about $\boldsymbol{X}^*$ such that $\lambda_i(\cdot)$ is continuously differentiable with a continuously varying associated eigenvector (Theorem 3.1.1 of [44]). Thus, each g_i is continuously differentiable at $\boldsymbol{X}^*$ and therefore regular at $\boldsymbol{X}^*$. The calculation of the Clarke subdifferential of g_i follows from the fact each eigenspace of $\boldsymbol{X}^*$ contains a single element, $\partial_C g_i(\boldsymbol{X}^*) = \left\{\boldsymbol{V}^*\mathrm{Diag}(\boldsymbol{a}_i)(\boldsymbol{V}^*)^\top\right\}$ where $\boldsymbol{X}^* = \boldsymbol{V}^*\mathrm{Diag}(\boldsymbol{\lambda}(\boldsymbol{X}^*))(\boldsymbol{V}^*)^\top$ is the eigendecomposition of $\boldsymbol{X}^*$. Hence,

$$\begin{aligned} &t_1\partial_C g_1(\boldsymbol{X}) + \cdots + t_p\partial_C g_p(\boldsymbol{X}) \\ &= \sum_{i=1}^{p} t_i \boldsymbol{V}^*\mathrm{Diag}(\boldsymbol{a}_i)(\boldsymbol{V}^*)^\top = \boldsymbol{V}^*\mathrm{Diag}\left(\sum_{i=1}^{p} t_i\boldsymbol{a}_i\right)(\boldsymbol{V}^*)^\top. \end{aligned} \tag{5}$$

To complete the proof, we prove the constraint qualification condition for Corollary 5.54 in [48] holds given our assumption on $\{\boldsymbol{a}_i \mid i \in \mathcal{I}(\boldsymbol{X}^*)\}$. Without loss of generality, we assume the first r constraints are active. Then the constraint qualification will hold provided $\mathbf{0} = t_1 \partial_C g_1(\boldsymbol{X}) + \cdots + t_r \partial_C g_r(\boldsymbol{X})$ if and only if $t_1 = \ldots = t_r = 0$. From (5), we see this is equivalent to $\mathbf{0} = \text{Diag}(t_1\boldsymbol{a}_1 + \cdots + t_r\boldsymbol{a}_r)$ if and only if $t_1 = \ldots = t_r = 0$ which follows from the assumptions. □

Comparing the necessary conditions with and without repeated eigenvalues inclines us to prefer the latter. The eigenspace dependence, the required regularity, and the constraint qualification limit the utility of Theorem 2 while such difficulties dissipate when the local minimizer has no repeated eigenvalues. In some cases, the constraint set ensures uniqueness; however, many constraints have feasible matrices with repeated eigenvalues. This seems unavoidable, but we now prove it is possible to approximate the original model and remove all feasible solutions with repeated eigenvalues. Letting $\epsilon > 0$, we define the approximated (SCO-Eig) model,

$$\begin{aligned} \min\ & F(\boldsymbol{X}) && \text{(SCO-Eig-}\epsilon\text{)} \\ \text{s.t. } & \boldsymbol{A}\lambda(\boldsymbol{X}) \leq \boldsymbol{b} \\ & \lambda_{i+1}(\boldsymbol{X}) \leq \lambda_i(\boldsymbol{X}) - \epsilon,\ i = 1, \ldots, n-1 \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n}. \end{aligned}$$

The next result shows that it is possible to bound the gap in the optimal values of the two models as a function of ϵ given certain assumptions.

Theorem 4 *Assume F is Lipschitz continuous with constant $L > 0$, the interior of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ is non-empty, and $[\boldsymbol{A}^\top | \boldsymbol{D}_n^\top] \in \mathbb{R}^{n\times(m+n-1)}$ has full rank. Let $\boldsymbol{X}^*$ be a global minimizer of (SCO-Eig). Then there exists $\epsilon_0 > 0$ such the constraint set of (SCO-Eig-ϵ) is non-empty for all $\epsilon \in [0, \epsilon_0]$, and, letting $\boldsymbol{X}_\epsilon^*$ be a global minimizer of (SCO-Eig-ϵ), we have for all $\epsilon \in [0, \epsilon_0]$*

$$|F(\boldsymbol{X}^*) - F(\boldsymbol{X}_\epsilon^*)| \leq L \cdot \chi([\boldsymbol{A}^\top | \boldsymbol{D}_n^\top]) \cdot (\sqrt{n-1})\epsilon,$$

where $\chi(\boldsymbol{Z}) := \left\{ \|\boldsymbol{Z}_{\mathcal{I}}^{-1}\|_2 \mid \mathcal{I} \subset [m+n-1],\ |\mathcal{I}| = n,\ \boldsymbol{Z}_{\mathcal{I}} \text{ non-singular} \right\}$[1] with $\boldsymbol{Z}_{\mathcal{I}}$ being the matrix formed from the columns of $\boldsymbol{Z}$ whose indices belong in $\mathcal{I}$.

Proof By our assumptions, there exists $\bar{\boldsymbol{X}} \in \mathcal{S}^{n\times n}$ such that $\boldsymbol{A}\boldsymbol{\lambda}(\bar{\boldsymbol{X}}) < \boldsymbol{b}$ and $\boldsymbol{D}_n\boldsymbol{\lambda}(\bar{\boldsymbol{X}}) < \mathbf{0}$. Therefore, there exists $\epsilon_0 > 0$ such that $\boldsymbol{A}\boldsymbol{\lambda}(\bar{\boldsymbol{X}}) \leq \boldsymbol{b}$ and $\boldsymbol{D}_n\boldsymbol{\lambda}(\bar{\boldsymbol{X}}) \leq -\epsilon_0\boldsymbol{e}$ which implies the constraint set of (SCO-Eig-ϵ) is non-empty for all $\epsilon \in [0, \epsilon_0]$.

Let $\boldsymbol{X}^* = \boldsymbol{V}^*\text{Diag}(\boldsymbol{\lambda}^*)(\boldsymbol{V}^*)^\top$ be a spectral decomposition of $\boldsymbol{X}^*$ and $P_{\mathcal{S}_\epsilon}(\boldsymbol{\lambda}^*)$ be the projection of $\boldsymbol{\lambda}^*$ onto $\mathcal{S}_\epsilon := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{x} \leq -\epsilon\boldsymbol{e}\}$. By the Lipschitz continuity of F, for any $\epsilon \in [0, \epsilon_0]$

$$\begin{aligned} |F(\boldsymbol{X}^*) - F(\boldsymbol{X}_\epsilon^*)| &= |F(\boldsymbol{V}^*\text{Diag}(\boldsymbol{\lambda}^*)(\boldsymbol{V}^*)^\top) - F(\boldsymbol{X}_\epsilon^*)| \\ &\leq |F(\boldsymbol{V}^*\text{Diag}(\boldsymbol{\lambda}^*)(\boldsymbol{V}^*)^\top) - F(\boldsymbol{V}^*\text{Diag}(P_{\mathcal{S}_\epsilon}(\boldsymbol{\lambda}^*))(\boldsymbol{V}^*)^\top)| \\ &\leq L\|\boldsymbol{V}^*\text{Diag}(\boldsymbol{\lambda}^*)(\boldsymbol{V}^*)^\top - \boldsymbol{V}^*\text{Diag}(P_{\mathcal{S}_\epsilon}(\boldsymbol{\lambda}^*))(\boldsymbol{V}^*)^\top\|_F \\ &= L\|\boldsymbol{\lambda}^* - P_{\mathcal{S}_\epsilon}(\boldsymbol{\lambda}^*)\|_2. \end{aligned} \tag{6}$$

[1] This constant was first presented by Dikin [13]. Equivalent definitions, as the one stated, were proven in [55, 61].

Using Hoffman's error bound for linear systems, we bound the distance between $\boldsymbol{\lambda}^*$ and $P_{\mathcal{S}_\epsilon}(\boldsymbol{\lambda}^*)$. By Theorem 3.6 in [61], for all $z \in \mathbb{R}^n$

$$\|z - P_{\mathcal{S}_\epsilon}(z)\|_2 \leq \chi([\boldsymbol{A}^\top | \boldsymbol{D}_n^\top]) \cdot \left\| \left([\boldsymbol{A}^\top | \boldsymbol{D}_n^\top]^\top z - [\boldsymbol{b}^\top | -\epsilon \boldsymbol{e}^\top]^\top \right)_+ \right\|_2 \tag{7}$$

where $(\boldsymbol{y})_+ := ((y_1)_+, \ldots, (y_n)_+)^\top$ for $\boldsymbol{y} \in \mathbb{R}^n$ with $(a)_+ := \max(0, a)$ for $a \in \mathbb{R}$ and for $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$ with full rank

$$\chi(\boldsymbol{Z}) := \left\{ \|\boldsymbol{Z}_{\mathcal{I}}^{-1}\|_2 \mid |\mathcal{I}| = m,\ \boldsymbol{Z}_{\mathcal{I}} \text{ is non-singular} \right\}, \tag{8}$$

where $\boldsymbol{Z}_{\mathcal{I}}$ denotes the submatrix matrix of $\boldsymbol{Z}$ composed of the columns of $\boldsymbol{Z}$ in the index set $\mathcal{I} \subset [n]$ [61]. Since $\boldsymbol{\lambda}^*$ is contained in $\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{x} \leq \boldsymbol{0}\}$,

$$\left\| \left([\boldsymbol{A}^\top | \boldsymbol{D}_n^\top]^\top \boldsymbol{\lambda}^* - [\boldsymbol{b}^\top | -\epsilon \boldsymbol{e}^\top]^\top \right)_+ \right\|_2 = \|(\boldsymbol{D}_n \boldsymbol{\lambda}^* + \epsilon \boldsymbol{e})_+\|_2 \leq \|\epsilon \boldsymbol{e}\|_2 = \epsilon\sqrt{n-1}. \tag{9}$$

Combining (6), (7) and (9) concludes the argument. □

4 Solving SCO-Eig with Linear Objective Functions

The simplest, non-trivial objective function to consider for (SCO-Eig) is a linear objective function,

$$\begin{aligned} &\min \langle \boldsymbol{C}, \boldsymbol{X} \rangle \\ &\text{s.t. } \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b} \\ &\quad \boldsymbol{X} \in \mathcal{S}^{n \times n}. \end{aligned} \tag{10}$$

As we saw in Sect. 2, the constraint set $\mathcal{S}_\lambda(\boldsymbol{A}, \boldsymbol{b})$ can be highly non-convex. Therefore, one may not expect to guarantee a global minimizer; however, Theorem 5 below shows a global minimizer is readily computable regardless of the constraint set. Moreover, the solution to (10) comes down to performing a single spectral decomposition and solving a single linear program.

Theorem 5 *Let* $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ *with* $\frac{1}{2}(\boldsymbol{C} + \boldsymbol{C}^\top) = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$ *for* $\boldsymbol{P} \in \mathcal{O}(n)$ *and* $\boldsymbol{\Omega} = \mathrm{Diag}([\omega_1, \omega_2, \ldots, \omega_n])$ *with* $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_n$. *Then a global minimizer of (10) is given by*

$$\boldsymbol{X}^* = \boldsymbol{P}\ \mathrm{Diag}([\lambda_n^*, \lambda_{n-1}^*, \cdots, \lambda_1^*])\boldsymbol{P}^\top \tag{11}$$

where

$$\boldsymbol{\lambda}^* \in \mathit{argmin} \left\{ \sum_{i=1}^n \omega_i \lambda_{n+1-i} \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0} \right\}. \tag{12}$$

Proof We may assume without loss of generality $\boldsymbol{C} \in \mathcal{S}^{n \times n}$. Observe,

$$\langle \boldsymbol{C}, \boldsymbol{X} \rangle = \left\langle \frac{1}{2}(\boldsymbol{C} + \boldsymbol{C}^\top), \boldsymbol{X} \right\rangle$$

for all $\boldsymbol{C} \in \mathbb{R}^{n \times n}$, so (10) can be rewritten such that the objective function is the inner product of two symmetric matrices. Thus, for the remainder of the proof, we assume $\boldsymbol{C} \in \mathcal{S}^{n \times n}$ and

$\boldsymbol{C} = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$ for orthogonal $\boldsymbol{P}$ and $\boldsymbol{\Omega} = \text{Diag}([\omega_1, \omega_2, \ldots, \omega_n])$ with $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_n$. Utilizing the spectral decomposition of $\boldsymbol{X}$,

$$\begin{aligned}
&\min\left\{\langle \boldsymbol{C}, \boldsymbol{X}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b},\ \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\} \\
&\quad = \min\left\{\langle \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top,\ \boldsymbol{Q}\,\text{Diag}(\boldsymbol{\lambda})\,\boldsymbol{Q}^\top\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0},\ \boldsymbol{Q} \in \mathcal{O}(n)\right\} \\
&\quad = \min\left\{\langle \boldsymbol{\Omega},\ \boldsymbol{P}^\top\boldsymbol{Q}\,\text{Diag}(\boldsymbol{\lambda})\,\boldsymbol{Q}^\top\boldsymbol{P}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0},\ \boldsymbol{Q} \in \mathcal{O}(n)\right\} \\
&\quad = \min\left\{\langle \boldsymbol{\Omega},\ \bar{\boldsymbol{Q}}^\top\,\text{Diag}(\boldsymbol{\lambda})\,\bar{\boldsymbol{Q}}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0},\ \bar{\boldsymbol{Q}} \in \mathcal{O}(n)\right\} \\
&\quad = \min\left\{\text{Tr}\left(\boldsymbol{\Omega}\,\bar{\boldsymbol{Q}}^\top\,\text{Diag}(\boldsymbol{\lambda})\,\bar{\boldsymbol{Q}}\right) \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0},\ \bar{\boldsymbol{Q}} \in \mathcal{O}(n)\right\}. \qquad (13)
\end{aligned}$$

By Theorem 2.1 in [40], it follows for any $\boldsymbol{\lambda} \in \mathbb{R}^n$ satisfying $\boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0}$ that

$$\min\left\{\text{Tr}\left(\boldsymbol{\Omega}\,\bar{\boldsymbol{Q}}^\top\,\text{Diag}(\boldsymbol{\lambda})\,\bar{\boldsymbol{Q}}\right) \mid \bar{\boldsymbol{Q}} \in \mathcal{O}(n)\right\} = \sum_{i=1}^{n} \omega_i \lambda_{n+1-i}.$$

Therefore,

$$\begin{aligned}
&\min\left\{\langle \boldsymbol{C}, \boldsymbol{X}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b},\ \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\} \\
&\quad = \min\left\{\sum_{i=1}^{n} \omega_i \lambda_{n+1-i} \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0}\right\}.
\end{aligned}$$

Let $\boldsymbol{\lambda}^* \in \text{argmin}\left\{\sum_{i=1}^{n} \omega_i \lambda_{n+1-i} \mid \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \leq \boldsymbol{0}\right\}$. In-order to compute $\boldsymbol{X}^*$ such that the constraints are satisfied and $\text{Tr}(\boldsymbol{C}^\top\boldsymbol{X}^*) = \sum_{i=1}^{n} \omega_i \lambda^*_{n+1-i}$, we must determine $\bar{\boldsymbol{Q}}^*$ such that,

$$\text{Tr}\left(\boldsymbol{\Omega}(\bar{\boldsymbol{Q}}^*)^\top\text{Diag}(\boldsymbol{\lambda}^*)\,\bar{\boldsymbol{Q}}^*\right) = \sum_{i=1}^{n} \omega_i \lambda^*_{n+1-i}. \qquad (14)$$

With $\bar{\boldsymbol{Q}}^*$ defined by (14) and using (13) to see $\bar{\boldsymbol{Q}} = \boldsymbol{Q}^\top\boldsymbol{P}$, a global minimizer $\boldsymbol{X}^*$ is given by,

$$\boldsymbol{X}^* = \boldsymbol{P}(\bar{\boldsymbol{Q}}^*)^\top\text{Diag}(\boldsymbol{\lambda}^*)\,\bar{\boldsymbol{Q}}^*\boldsymbol{P}^\top.$$

Since $\bar{\boldsymbol{Q}}^* = [\boldsymbol{e}_n\ \boldsymbol{e}_{n-1}\ldots\boldsymbol{e}_1]$ solves (14), we obtain our final result. □

Remark 2 Though Theorem 5 was stated and proven for $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{X} \in \mathcal{S}^{n\times n}$, the same general argument applies with slight modifications if $\boldsymbol{X} \in \mathbb{C}^{n\times n}$ is Hermitian and $\boldsymbol{C} \in \mathbb{C}^{n\times n}$.

Theorem 5 provides a straight-forward procedure for computing a global minimizer to (10) by solving a single linear program and performing one spectral decomposition, and the result was independent of the convexity of $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$; therefore, (10) is solvable in polynomial-time and constitutes a reasonable subproblem for an algorithm.

Upon completing the paper, we were made aware of a relatively recent and interesting though, unfortunately, unnoticed paper by Gowda [19]. This paper presents results of an equivalent nature to the results presented in this section and the next; however, the author uses a more technical approach than our own which is more direct and tailored to our goals of developing algorithms to solve (SCO-Eig). The importance of this result lies in leveraging it toward algorithm design. We present in this paper the first algorithms for general spectral optimization taking advantage of these tractable sub-procedures.

5 Projecting onto the Eigenvalue Constraint Set

Projecting onto the constraint set is a crucial procedure in numerous constrained optimization algorithms. In this section, we utilize an argument similar to the one applied in Sect. 4 to compute projections onto $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$. That is, given $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$, we solve

$$\begin{aligned} \min\ & \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \\ \text{s.t. } & \boldsymbol{A}\lambda(\boldsymbol{X}) \le \boldsymbol{b} \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n}. \end{aligned} \tag{15}$$

Theorem 6 *Let* $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ *with* $\frac{1}{2}(\boldsymbol{Y}+\boldsymbol{Y}^\top) = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$ *for* $\boldsymbol{P} \in \mathcal{O}(n)$ *and* $\boldsymbol{\Omega} = \mathrm{Diag}([\omega_1, \omega_2, \ldots, \omega_n])$ *with* $\omega_1 \le \omega_2 \le \ldots \le \omega_n$ *and* $\bar{\boldsymbol{\omega}} := [\omega_n, \omega_{n-1}, \ldots, \omega_1]^\top$. *Then*

$$\boldsymbol{X}^* = \boldsymbol{P}\,\mathrm{Diag}([\lambda_n^*, \lambda_{n-1}^*, \ldots, \lambda_1^*])\boldsymbol{P}^\top$$

is a global minimizer of (15) where

$$\boldsymbol{\lambda}^* \in \mathit{argmin}\left\{\frac{1}{2}\|\boldsymbol{\lambda}-\bar{\boldsymbol{\omega}}\|_2^2 \mid \boldsymbol{A}\boldsymbol{\lambda} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \le \boldsymbol{0}\right\}. \tag{16}$$

Proof Let $\frac{1}{2}(\boldsymbol{Y}+\boldsymbol{Y}^\top) = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$ with $\boldsymbol{P} \in \mathcal{O}(n)$ and $\boldsymbol{\Omega} = \mathrm{Diag}([\omega_1, \omega_2, \ldots, \omega_n])$ such that $\omega_1 \le \omega_2 \le \ldots \le \omega_n$. Define $\bar{\boldsymbol{\Omega}} := -\boldsymbol{\Omega}$ and $\bar{\boldsymbol{\omega}} := [\omega_n, \omega_{n-1}, \ldots, \omega_1]^\top$. Utilizing the spectral decomposition of $\boldsymbol{X}$,

$$\begin{aligned} &\min\left\{\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \mid \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \le \boldsymbol{b}, \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\} \\ &= \min\left\{\frac{1}{2}\|\boldsymbol{X}\|_F^2 - \langle \boldsymbol{X}, \boldsymbol{Y}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \le \boldsymbol{b}, \boldsymbol{X} \in \mathcal{S}^{n\times n}\right\} + \frac{1}{2}\|\boldsymbol{Y}\|_F^2 \\ &= \min\left\{\frac{1}{2}\|\boldsymbol{\lambda}\|^2 - \langle \boldsymbol{Q}\,\mathrm{Diag}(\boldsymbol{\lambda})\,\boldsymbol{Q}^\top, \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \le \boldsymbol{0},\ \boldsymbol{Q} \in \mathcal{O}(n)\right\} + \frac{1}{2}\|\boldsymbol{Y}\|_F^2 \\ &= \min\left\{\frac{1}{2}\|\boldsymbol{\lambda}\|^2 + \langle \bar{\boldsymbol{Q}}^\top\,\mathrm{Diag}(\boldsymbol{\lambda})\,\bar{\boldsymbol{Q}}, \bar{\boldsymbol{\Omega}}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \le \boldsymbol{0},\ \bar{\boldsymbol{Q}} \in \mathcal{O}(n)\right\} + \frac{1}{2}\|\boldsymbol{Y}\|_F^2 \\ &= \min\left\{\frac{1}{2}\|\boldsymbol{\lambda}\|^2 - \langle \bar{\boldsymbol{\omega}}, \boldsymbol{\lambda}\rangle \mid \boldsymbol{A}\boldsymbol{\lambda} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \le \boldsymbol{0}\right\} + \frac{1}{2}\|\boldsymbol{Y}\|_F^2 \\ &= \min\left\{\frac{1}{2}\|\boldsymbol{\lambda}-\bar{\boldsymbol{\omega}}\|_2^2 \mid \boldsymbol{A}\boldsymbol{\lambda} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{\lambda} \le \boldsymbol{0}\right\} + \frac{1}{2}(\|\boldsymbol{Y}\|_F^2 - \|\bar{\boldsymbol{\omega}}\|_2^2) \\ &= \frac{1}{2}\left(\|\mathrm{Proj}_{\mathcal{C}}(\bar{\boldsymbol{\omega}}) - \bar{\boldsymbol{\omega}}\|_2^2 + \|\boldsymbol{Y}\|_F^2 - \|\bar{\boldsymbol{\omega}}\|_2^2\right) \end{aligned} \tag{17}$$

where $\mathcal{C} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{x} \le \boldsymbol{0}\}$ and $\mathrm{Proj}_{\mathcal{C}}(\cdot)$ is the projection operator onto $\mathcal{C}$. Note, the fourth equality above follows from Theorem 2.1 in [40]. Let $\boldsymbol{\lambda}^* := \mathrm{Proj}_{\mathcal{C}}(\bar{\boldsymbol{\omega}})$. Note, if $\bar{\boldsymbol{Q}}^* = [\boldsymbol{e}_n\ \boldsymbol{e}_{n-1}\ \ldots\ \boldsymbol{e}_1]$, then

$$\mathrm{Tr}(\bar{\boldsymbol{\Omega}}(\bar{\boldsymbol{Q}}^*)^\top\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\,\bar{\boldsymbol{Q}}^*) = \frac{1}{2}\left(\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\omega}}\|^2 - \|\boldsymbol{\lambda}^*\|^2 - \|\bar{\boldsymbol{\omega}}\|^2\right) = \langle \boldsymbol{\lambda}^*, -\bar{\boldsymbol{\omega}}\rangle$$

and by the change-of-variable above $\boldsymbol{X}^* = \boldsymbol{P}(\bar{\boldsymbol{Q}}^*)^\top\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\,\bar{\boldsymbol{Q}}^*\boldsymbol{P}^\top$ is a global minimizer of (15). □

Thus, similar to the linear objective problem, projecting onto the eigenvalue constraint only requires computing a spectral decomposition and solving a convex optimization problem. In this case, a projection onto $\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{D}_n\boldsymbol{x} \leq \boldsymbol{0}\}$ must be computed. This problem can be solved via any convex optimization solver, but specialized fast algorithms for projecting onto a polyhedron have been developed [20]. We now utilize our results to develop two optimization algorithms which obtain first-order stationary points to (SCO-Eig) when the constraint set is convex.

6 Spectrally Constrained Solvers

We develop two first-order algorithms to compute stationary points for (SCO-Eig). Our capacity to solve the linear objective problem motivates the construction of a Frank-Wolfe algorithm and accessible projections make possible the development of a projected gradient method.

6.1 Inexact Projected Gradient Method

For $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ let $\text{Proj}_{\mathcal{S}_\lambda}(\cdot)$ be the operator which maps $\boldsymbol{Y}$ to an element of the argmin set of (15), i.e.,

$$\text{Proj}_{\mathcal{S}_\lambda}(\boldsymbol{Y}) \in \text{argmin}\left\{\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \mid \boldsymbol{X} \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})\right\}.$$

For $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ convex, the projection is unique. If the set is non-convex multiple optimal projections are likely present. Since inexact computations are the reality in implementation, we introduce a notion of inexact projections.

Definition 1 For convex $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ and parameter $\delta \geq 0$, the set of inexact projections of $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ onto $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ is

$$\text{P}^{\delta}_{\mathcal{S}_\lambda}(\boldsymbol{Y}) := \{\boldsymbol{Z} \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b}) \mid \langle \boldsymbol{Z}-\boldsymbol{Y}, \boldsymbol{X}-\boldsymbol{Z}\rangle \geq -\delta,\ \forall \boldsymbol{X} \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})\}. \tag{18}$$

The standard results for projections onto convex sets tell us $\boldsymbol{Z}^*$ is the optimal projection of $\boldsymbol{Y}$ onto convex $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ if and only if

$$\langle \boldsymbol{Z}^*-\boldsymbol{Y}, \boldsymbol{X}-\boldsymbol{Z}^*\rangle \geq 0,\ \forall \boldsymbol{X} \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b}).$$

Therefore, the set of inexact projections are the points which inexactly satisfy this condition where the level of inexactness is controlled by δ. Clearly, $\text{P}^0_{\mathcal{S}_\lambda}(\boldsymbol{Y}) = \text{Proj}_{\mathcal{S}_\lambda}(\boldsymbol{Y})$ when $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ is a convex set.

For the sake of our analysis, we make the following assumption throughout Sect. 6

Assumption 1 $\mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ is convex.

This assumption enables us to have a clear notion of first-order ϵ-stationary points for (SCO-Eig).

Definition 2 Let $\epsilon \geq 0$. The point $\boldsymbol{X}^* \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})$ is a first-order ϵ-stationary point of (SCO-Eig) provided

$$\min\left\{\langle \nabla F(\boldsymbol{X}^*), \boldsymbol{X}-\boldsymbol{X}^*\rangle \mid \|\boldsymbol{X}-\boldsymbol{X}^*\|_F \leq 1,\ \boldsymbol{X} \in \mathcal{S}_\lambda(\boldsymbol{A},\boldsymbol{b})\right\} \geq -\epsilon.$$

Algorithm 1 presents our inexact projected gradient method for solving (SCO-Eig). The method can be neatly described. A point $\boldsymbol{X}$ is selected from $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$ and the traditional gradient update step is taken to compute $\boldsymbol{X}_+ = \boldsymbol{X} - \alpha \nabla F(\boldsymbol{X})$ where $\alpha > 0$ is the stepsize. Since it is not necessarily true $\boldsymbol{X}_+ \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$, because α could be too long of a step or $\nabla F(\boldsymbol{X})$ might point out of the constraint set, feasibility is maintained by projecting $\boldsymbol{X}_+$ onto the constraint. To save computational expenses, the projection is done inexactly. Line-search is performed at each iteration to ensure a sufficient decrease is obtained. The algorithm terminates when the gap between consecutive iterates decreases below a provided tolerance.

The approach is a fairly straightforward implementation of the traditional method in constrained optimization. The key novelty of Algorithm 1 comes from the fact we can solve the projection problem by Theorem 6. Without this result the algorithm would not be implementable.

Algorithm 1 Inexact Projected Gradient Method

Require: $\boldsymbol{X}_0 \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$; $\epsilon > 0$; $\delta \in [0, 1)$; $\alpha > 0$; $\tau_1 \in (0, 1)$; $h > 0$
for $k = 0, 1, 2 \ldots$ **do**
  $h_k = h$
  $\boldsymbol{X}_{k+1} = \mathrm{P}^{\delta}_{\mathcal{S}_{\boldsymbol{\lambda}}}(\boldsymbol{X}_k - h_k \nabla F(\boldsymbol{X}_k))$
  if $\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\|_F \leq \epsilon$ **then**
    Return $\boldsymbol{X}_k$
  else
    while $F(\boldsymbol{X}_{k+1}) > F(\boldsymbol{X}_k) - \alpha \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2$ **do**
      $h_k = \tau_1 \cdot h_k$
      $\boldsymbol{X}_{k+1} = \mathrm{P}^{\delta}_{\mathcal{S}_{\boldsymbol{\lambda}}}(\boldsymbol{X}_k - h_k \nabla F(\boldsymbol{X}_k))$
    end while
  end if
end for

We now explicitly state the convergence result of Algorithm 1. A more generalized version of the algorithm is proven in Sect. 7 which is applicable to the block model described in Remark 3.

Theorem 7 *Assume F is gradient Lipschitz with parameter $L > 0$, the initial level set, i.e., $\{\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) \mid F(\boldsymbol{X}) \leq F(\boldsymbol{X}_0)\}$, is a bounded subset of $\mathcal{S}^{n \times n}$ with diameter D, there exists F^* such that $F^* \leq F(\boldsymbol{X})$ for all $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$, there exists $M > 0$ such that $\|\nabla F(\boldsymbol{X}_k)\|_F \leq M$ for all $k \geq 0$, and Assumption 1 holds. If inexact projections are computed with sufficient accuracy, dependent on ϵ, then Algorithm 1 will converge to a first-order ϵ-stationary point of (SCO-Eig) in $\mathcal{O}(\epsilon^{-2})$ iterations. More explicitly, a first-order ϵ-stationary point will be returned after no more than*

$$\left(\frac{4(D + Mh_{low} + 1)^2}{h_{low}^2}\right)\left(\frac{F(\boldsymbol{X}_0) - F^*}{\alpha}\right)\frac{1}{\epsilon^2} \text{ iterations}$$

provided the accuracy of the inexact projections, δ, satisfies $\delta \leq \min\left\{\frac{h_{low}}{2}\epsilon, \frac{1}{2}\epsilon_{tol}^2\right\}$, where $h_{low} := \tau_1/(L + 2\alpha)$ and $\epsilon_{tol} := \frac{h_{low}\epsilon}{2(D + Mh_{low} + 1)}$.

6.2 Inexact Frank–Wolfe Algorithm

We now present a Frank-Wolfe algorithm for solving (SCO-Eig). The algorithm and analysis we present are an extension of the work in the concise technical report of Lacoste-Julien [30]. The author in this paper presents the first Frank-Wolfe method for smooth non-convex functions over a compact and convex constraint. Our approach extends the work in [30] by removing the compactness assumption. To accomplish this, we utilized a different notion of first-order stationarity, replaced the compactness assumption with the weaker assumption that the initial level set of the objective function is bounded, and introduced a modified subproblem for our model. Traditional Frank-Wolfe approaches would require solving subproblems of the form

$$\begin{aligned} \min\ & \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \\ \text{s.t. } & \boldsymbol{A}\boldsymbol{\lambda}(\boldsymbol{X}) \leq \boldsymbol{b} \\ & \|\boldsymbol{X} - \boldsymbol{X}_0\|_F \leq 1 \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n}. \end{aligned} \tag{19}$$

This model appears outside the scope of (SCO-Eig); however, Theorem 8 below shows (19) can be bounded by an alternative optimization model which only contains linear constraints on the eigenvalues.

Theorem 8 *Given any $\boldsymbol{X}_0 \in \mathcal{S}^{n\times n}$, we have the following upper bound to (19)*

$$\begin{aligned} &\left| \min \{ \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{X} - \boldsymbol{X}_0\|_F \leq 1 \} \right| \\ &\leq \left| \min \{ \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq 1 \} \right|. \end{aligned} \tag{20}$$

Proof From Section 6.3 of [24], we know for any $\boldsymbol{X}, \boldsymbol{X}_0 \in \mathcal{S}^{n\times n}$

$$\|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_2 \leq \|\boldsymbol{X} - \boldsymbol{X}_0\|_2 \text{ implying } \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq \|\boldsymbol{X} - \boldsymbol{X}_0\|_F.$$

Thus, given any $\boldsymbol{X} \in \mathcal{S}^{n\times n}$ such that $\|\boldsymbol{X} - \boldsymbol{X}_0\|_F \leq 1$ it follows $\|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq 1$ and

$$\left\{ \boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \|\boldsymbol{X} - \boldsymbol{X}_0\|_F \leq 1 \right\} \subset \left\{ \boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq 1 \right\}.$$

Hence

$$\begin{aligned} &\min \{ \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{X} - \boldsymbol{X}_0\|_F \leq 1 \} \\ &\geq \min \{ \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq 1 \}, \end{aligned} \tag{21}$$

and we know the optimal value of both models is non-positive since $\boldsymbol{X}_0$ is a feasible solution for each model. Therefore, the right-hand side of (21) is greater in terms of absolute magnitude. □

Theorem 8 results in our ability to approximate the solution to (19) by solving an alternative optimization model which only appends additional linear inequality constraints on the eigenvalues. Since Definition 2 is equivalent to the left-hand side of (20) being less than ϵ, it follows the right-hand side of (20) provides an estimate of the optimality of $\boldsymbol{X}_0$ via a tractable subproblem. Therefore, in our Frank-Wolfe approach, instead of solving (19), we solve models of the form

$$\min \{ \langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_0 \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{X}_0)\|_\infty \leq 1 \}$$

and use these solutions to bound the first-order stationarity condition of (SCO-Eig).

Algorithm 2 is our proposed inexact Frank-Wolfe algorithm for (SCO-Eig). At each iteration it seeks to improve upon the current iterate by minimizing a first-order approximation of the function over the constraint set. We approximate the traditional Frank-Wolfe subproblem by Theorem 8. The subproblem then produces a direction which will decrease the value of the objective function, and a stepsize is determined which maintains the feasibility of the next iterate.

Algorithm 2 Inexact Frank-Wolfe algorithm

Require: $\boldsymbol{X}_0 \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$; $\epsilon > 0$; $\alpha \in [0, 1)$; $\delta \in [0, \alpha\epsilon)$; $\Theta > 0$

for $k = 0, 1, 2, \ldots$ **do**

Approximately compute a solution, $\boldsymbol{D}_k$, to

$$m_k^* := \min \{\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{D} - \boldsymbol{X}_k \rangle \mid \boldsymbol{D} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{\lambda}(\boldsymbol{D}) - \boldsymbol{\lambda}(\boldsymbol{X}_k)\|_\infty \leq 1\} \tag{22}$$

such that $m_k := \langle \nabla F(\boldsymbol{X}_k), \boldsymbol{D}_k - \boldsymbol{X}_k \rangle \leq 0$ and is within δ of m_k^*

if $|m_k| \leq (1 - \alpha)\epsilon$ **then**

Return $\boldsymbol{X}_k$

else

Set $\gamma_k := \min \{|m_k|/\Theta, 1\}$

Update $\boldsymbol{X}_{k+1} := \boldsymbol{X}_k + \gamma_k (\boldsymbol{D}_k - \boldsymbol{X}_k)$

end if

end for

Theorem 9 states the sublinear convergence rate of Algorithm 2 to stationary points of (SCO-Eig). The convergence analysis of the algorithm is provided in the next section.

Theorem 9 *Assume F is gradient Lipschitz with parameter $L > 0$, the initial level set is bounded, there exists F^* such that $F^* \leq F(\boldsymbol{X})$ for all $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b})$, and Assumption 1 holds. Define*

$$\rho := \sup \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 | F(\boldsymbol{X}) \leq F(\boldsymbol{X}_0), \|\boldsymbol{\lambda}(\boldsymbol{X}) - \boldsymbol{\lambda}(\boldsymbol{Y})\|_\infty \leq 1, \boldsymbol{Y}, \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}) \right\}.$$

If $\Theta \geq \rho L$, then Algorithm 2 will compute a first-order ϵ-stationary point to (SCO-Eig) in $\mathcal{O}(\epsilon^{-2})$ iterations. More precisely, a first-order ϵ-stationary point will be obtained after no more than

$$K \geq \left\lceil \frac{\max \{2(F(\boldsymbol{X}_0) - F^*), \Theta\}^2}{(1 - \alpha)^2 \epsilon^2} - 1 \right\rceil \text{ iterations.}$$

The condition in Theorem 9 requiring $\Theta \geq \rho L$ might appear unruly since both ρ and L could be unknown. In practice this difficulty is overcome by updating Θ on an iterative basis. For example, if $\Theta \geq \rho L$, then the inequality in (26) shall hold. If this inequality fails, then Θ can be increased until the inequality is satisfied at the current iterate. In our implementation of Algorithm 2, we update the value of Θ at each iteration depending on whether or not the value of the objective function was improved. This is why we ensure an improvement of the value of the objective function.

Both Algorithms 1 and 2 converge sublinearly to first-order stationary points of (SCO-Eig) under reasonable assumptions. The convexity assumption is necessary for the current convergence analysis of both algorithms and is required to maintain the feasibility of the iterates generated by Algorithm 2. The projected gradient method however can be applied successfully to instances of (SCO-Eig) with non-convex constraints and maintain feasibility. We

observe this, for example, in Sect. 8.2 when Algorithm 1 is applied to solve systems of quadratic equations.

Remark 3 The theory and algorithms we have developed extend to the block optimization model

$$\begin{aligned} &\min\ F(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \\ &\text{s.t. } \boldsymbol{X}_i \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}_i, \boldsymbol{b}_i),\ i = 1, \ldots, k. \end{aligned} \tag{23}$$

Letting $\mathcal{V} := \mathcal{S}^{n_1 \times n_1} \times \cdots \times \mathcal{S}^{n_k \times n_k}$ with vectors $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ and associated inner product $\langle \mathcal{X}, \mathcal{Y} \rangle_{\mathcal{V}} := \sum_{i=1}^{k} \langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$, the analysis in Sect. 7 proves Algorithm 1 computes first-order stationary points of (23) where exact projections onto $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}_1, \boldsymbol{b}_1) \times \cdots \times \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}_k, \boldsymbol{b}_k)$ are readily computed by projecting onto each component $\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}_i, \boldsymbol{b}_i)$. Similarly, the required subproblem to implement a modified version of Algorithm 2 on (23) is equivalent to solving k subproblems of the form described in Sect. 6.2. Proving convergence of a modified version of Algorithm 2 to stationary points of (23) only requires minor alterations to the analysis presented in the next section.

7 Convergence Analysis

7.1 Proof of Theorem 9

Proof We first show ρ provides a usable bound for all iterations of Algorithm 2. If $|m_0| \leq \Theta$, then by the gradient Lipschitz condition on F we have,

$$\begin{aligned} F(\boldsymbol{X}_1) &\leq F(\boldsymbol{X}_0) + \gamma_0 \langle \nabla F(\boldsymbol{X}_0), \boldsymbol{D}_0 - \boldsymbol{X}_0 \rangle + \frac{L\gamma_0^2}{2} \|\boldsymbol{D}_0 - \boldsymbol{X}_0\|_F^2 \\ &\leq F(\boldsymbol{X}_0) + \gamma_0 \langle \nabla F(\boldsymbol{X}_0), \boldsymbol{D}_0 - \boldsymbol{X}_0 \rangle + \frac{\rho L \gamma_0^2}{2} \\ &\leq F(\boldsymbol{X}_0) + \gamma_0 m_0 + \frac{\Theta}{2}\gamma_0^2 \\ &\leq F(\boldsymbol{X}_0) - \frac{|m_0|^2}{\Theta} + \frac{1}{2\Theta}|m_0|^2 \\ &= F(\boldsymbol{X}_0) - \left(\frac{1}{2\Theta}\right)|m_0|^2 \end{aligned} \tag{24}$$

where the second inequality follows from the definition of ρ. If instead we have $|m_0| > \Theta$ then

$$F(\boldsymbol{X}_1) \leq F(\boldsymbol{X}_0) - |m_0| + \frac{\Theta}{2} < F(\boldsymbol{X}_0) - \frac{\Theta}{2}. \tag{25}$$

Therefore, regardless of how the stepsize is computed, $F(\boldsymbol{X}_1) \leq F(\boldsymbol{X}_0)$ which implies $\boldsymbol{X}_1$ is contained in the initial level set. So, the bound provided by ρ remains valid for all iterations and by induction on the iteration count we have

$$F(\boldsymbol{X}_{k+1}) \leq F(\boldsymbol{X}_k) - \min\left\{\frac{|m_k|^2}{2\Theta}, |m_k| - \frac{\Theta}{2}\mathbf{1}_{\{|m_k| > \Theta\}}\right\}. \tag{26}$$

Thus, using the same argument presented for Theorem 1 in [30] it follows for all $K \geq 0$,

$$\min_{0 \leq k \leq K} |m_k| \leq \frac{\max\{2(F(\boldsymbol{X}_0) - F^*), \Theta\}}{\sqrt{K+1}}.$$

Since m_k measures the first-order stationary condition inexactly, it then follows an ϵ-stationary point will be obtained provided $|m_t| \leq (1-\alpha)\epsilon$ where $t \in \{0, 1 \ldots, K\}$ produces the minimum value of $|m_k|$ over all the iterates since $|m_t^*| \leq |m_t| \leq |m_t| + \delta \leq \epsilon$. Finally, by (20) it follows if $|m_t^*| \leq \epsilon$ then

$$\left| \min \{\langle \boldsymbol{C}, \boldsymbol{X} - \boldsymbol{X}_t \rangle \mid \boldsymbol{X} \in \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{b}),\ \|\boldsymbol{X} - \boldsymbol{X}_t\|_F \leq 1\} \right| \leq \epsilon.$$

□

7.2 Proof of Theorem 7

In this section, we state and prove a general convergence result for Algorithm 1. Our argument is given for a general vector space $\mathcal{V}$ with associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and induced norm $\|\cdot\|_{\mathcal{V}}$. For the remainder of this section, we drop the subscript and refer to the inner product and norm on $\mathcal{V}$ as $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ respectively. Thus, the Frobenius norm used in the description of Algorithm 1 has been replaced with our general norm for $\mathcal{V}$. We assume the norm associated with $\mathcal{V}$ defines the first-order ϵ-stationary condition. Let P_C^{δ} denote the inexact projection onto a convex subset $\mathcal{C}$ of $\mathcal{V}$.

Theorem 10 *Assume F is gradient Lipschitz with parameter $L > 0$, the initial level set, i.e., $\{\boldsymbol{X} \in \mathcal{C} \mid F(\boldsymbol{X}) \leq F(\boldsymbol{X}_0)\}$, is a bounded subset of $\mathcal{V}$ with diameter D, there exists F^* such that $F^* \leq F(\boldsymbol{X})$ for all $\boldsymbol{X} \in \mathcal{C}$, there exists $M > 0$ such that $\|\nabla F(\boldsymbol{X}_k)\| \leq M$ for all $k \geq 0$, and $\mathcal{C}$ is a convex subset of the normed vector space $\mathcal{V}$. If inexact projections are computed with sufficient accuracy, dependent on ϵ, then Algorithm 1 will converge to a first-order ϵ-stationary point of* $\min\{F(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{C}\}$ *in $\mathcal{O}(\epsilon^{-2})$ iterations. More explicitly, a first-order ϵ-stationary point will be returned after no more than*

$$\left(\frac{4(D + Mh_{low} + 1)^2}{h_{low}^2}\right)\left(\frac{F(\boldsymbol{X}_0) - F^*}{\alpha}\right)\frac{1}{\epsilon^2} \text{ iterations}$$

provided the accuracy of the inexact projections, δ, satisfies $\delta \leq \min\left\{\frac{h_{low}}{2}\epsilon, \frac{1}{2}\epsilon_{tol}^2\right\}$, where $h_{low} := \tau_1/(L + 2\alpha)$ and $\epsilon_{tol} := \frac{h_{low}\epsilon}{2(D+Mh_{low}+1)}$.

Proof Compute $\boldsymbol{X}_{k+1} \in \mathrm{P}_{\mathcal{C}}^{\delta}(\boldsymbol{X}_k - h_k \nabla F(\boldsymbol{X}_k))$. If $\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \leq \epsilon_{\text{tol}}$, we shall prove $\boldsymbol{X}_k$ is a first-order ϵ-stationary point. If $\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| > \epsilon_{\text{tol}}$, we will show a sufficient decrease can be obtained provided the stepsize h_k and precision $\delta > 0$ are appropriately sized. To these ends, assume $\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| > \epsilon_{\text{tol}}$. If $h_k \leq (L + 2\alpha)^{-1}$ and $\delta \leq \frac{1}{2}\epsilon_{\text{tol}}^2$, then it follows

$$\begin{aligned}
&\langle \boldsymbol{X}_k - h_k \nabla F(\boldsymbol{X}_k) - \boldsymbol{X}_{k+1}, \boldsymbol{X}_k - \boldsymbol{X}_{k+1} \rangle \leq \delta \\
\implies &\langle h_k \nabla F(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k \rangle \leq \delta - \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2 \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k \rangle \leq h_k^{-1}\left(\delta - \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2\right).
\end{aligned}$$

By our assumptions, $\delta \leq \frac{1}{2}\epsilon_{\text{tol}}^2 \leq \frac{1}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2$. Therefore,

$$\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k \rangle \leq \frac{-h_k^{-1}}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2 \leq -\left(\frac{L + 2\alpha}{2}\right)\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2.$$

Thus, from the gradient Lipschitz assumption on F and the above inequality,

$$F(\boldsymbol{X}_{k+1}) \leq F(\boldsymbol{X}_k) + \langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k \rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2$$

$$\leq F(\boldsymbol{X}_k) - \left(-\frac{L}{2} + \alpha + \frac{L}{2}\right)\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2$$
$$= F(\boldsymbol{X}_k) - \alpha\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2. \tag{27}$$

Hence, the line-search shall terminate with a sufficient decrease obtained after no more than $\lceil \log_2(h(2\alpha + L))\rceil$ inner iterations. So, for all $k \geq 0$ we have

$$\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2 \leq \frac{F(\boldsymbol{X}_k) - F(\boldsymbol{X}_{k+1})}{\alpha}.$$

Summing this inequality up from $k = 0$ to $k = K - 1$ and using the lower bound F^* we obtain

$$\min_{0\leq k\leq K-1} \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|^2 \leq \left(\frac{F(\boldsymbol{X}_0) - F^*}{\alpha}\right)\frac{1}{K}. \tag{28}$$

We now prove if consecutive iterates generated by Algorithm 1 are sufficiently close together and the approximated projections are sufficiently accurate then a first-order ϵ-stationary point has been obtained. By the Lipschitz gradient assumption on F and the boundedness of the initial level set there exists $M > 0$ such that $\|\nabla F(\boldsymbol{X}_k)\| \leq M$ for all $k \geq 0$. Let D be the diameter of the initial level set where the diameter of a set S is defined as $\sup\{\|\boldsymbol{x} - \boldsymbol{y}\| \mid \boldsymbol{x}, \boldsymbol{y} \in S\}$. D is finite due to the boundedness of the initial level. Also, since a sufficient decrease is obtained if $h_k \leq (L + 2\alpha)^{-1}$, it follows $h_{\text{low}} = \tau_1(L + 2\alpha)^{-1}$ lower bounds the stepsize h_k for all $k \geq 0$. Assume $\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\| \leq \epsilon_{\text{tol}}$ and $\delta \leq \min\left\{\frac{h_{\text{low}}}{2}\epsilon,\ \frac{1}{2}\epsilon_{\text{tol}}^2\right\}$, where

$$\epsilon_{\text{tol}} := \frac{h_{\text{low}}\epsilon}{2(D + Mh_{\text{low}} + 1)}.$$

Then for all $\boldsymbol{X} \in \mathcal{C}$ such that $\|\boldsymbol{X} - \boldsymbol{X}_k\| \leq 1$ we have

$$\begin{aligned}
&\langle \boldsymbol{X}_k - h_k\nabla F(\boldsymbol{X}_k) - \boldsymbol{X}_{k+1},\ \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \leq \delta \\
\implies &\langle -h_k\nabla F(\boldsymbol{X}_k),\ \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \leq \delta + \langle \boldsymbol{X}_{k+1} - \boldsymbol{X}_k, \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \\
\implies &\langle -h_k\nabla F(\boldsymbol{X}_k),\ \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \leq \\
&\qquad \delta + \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \cdot \|\boldsymbol{X} - \boldsymbol{X}_k + \boldsymbol{X}_k - \boldsymbol{X}_{k+1}\| \\
\implies &\langle -h_k\nabla F(\boldsymbol{X}_k),\ \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \leq \\
&\qquad \delta + \|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \cdot (\|\boldsymbol{X} - \boldsymbol{X}_k\| + \|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\|) \\
\implies &\langle -h_k\nabla F(\boldsymbol{X}_k),\ \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \leq \delta + (1 + D)\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \geq -\delta h_k^{-1} - (1 + D)h_k^{-1}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_{k+1}\rangle \geq -\delta h_{\text{low}}^{-1} - (1 + D)h_{\text{low}}^{-1}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_k\rangle \geq \\
&\qquad -\delta h_{\text{low}}^{-1} - (1 + D)h_{\text{low}}^{-1}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| - \langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X}_k - \boldsymbol{X}_{k+1}\rangle \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_k\rangle \geq \\
&\qquad -\delta h_{\text{low}}^{-1} - (1 + D)h_{\text{low}}^{-1}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| - M\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\| \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_k\rangle \geq -\delta h_{\text{low}}^{-1} - ((1 + D)h_{\text{low}}^{-1} + M)\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\| \\
\implies &\langle \nabla F(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_k\rangle \geq -\epsilon
\end{aligned} \tag{29}$$

where the first implication comes from the definition of $\boldsymbol{X}_{k+1} \in \mathrm{P}_{\mathcal{C}}^{\delta}(\boldsymbol{X}_k - h_k\nabla F(\boldsymbol{X}_k))$, the sixth comes from h_{low} lower bounding h_k for all k, the eighth is a product of the bound

on the norm of the gradient of F, and the last implication follows from the bounds on δ and $\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\|$. Therefore, under these conditions $\boldsymbol{X}_k$ is a first-order ϵ-stationary point, and from (28) a point satisfying $\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\| \leq \epsilon_{\text{tol}}$ will be obtained within K iterations provided

$$K \geq \left(\frac{4(D + Mh_{\text{low}} + 1)^2}{h_{\text{low}}^2}\right)\left(\frac{F(\boldsymbol{X}_0) - F^*}{\alpha}\right) \cdot \frac{1}{\epsilon^2}.$$

□

8 Numerical Experiments

The (SCO-Eig) paradigm applies to many constrained matrix problems: semidefinite programming [53], condition number constraints [52, 58], and rank constrained optimization [39, 62] to list a few. In this section, we first demonstrate the performance of Algorithms 1 and 2 on a popular preconditioning model to showcase how the convergence theory aligns with what is observed in practice. Then, we demonstrate the applicability of our methodology on two important tasks: solving systems of quadratic equations and matrix completion. Our numerical results demonstrate our method can outperform classical approaches for solving quadratic systems, such as Newton's method, and best classical approaches to matrix completion when additional spectral information is available.

8.1 Preconditioning

One could argue solving systems of linear equations is the most important task in applied mathematics. Many iterative methods have been devised to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ full rank and $\boldsymbol{b} \in \mathbb{R}^m$ such as the Jacobi method [50] and the conjugate gradient method [22]. Methods for solving linear systems often converge linearly with the rate dependent upon the condition number of the matrix $\boldsymbol{A}$, i.e.,

$$\kappa(\boldsymbol{A}) := \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$$

where $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ are the largest and smallest singular values of $\boldsymbol{A}$ respectively. If $\boldsymbol{A} \in \mathcal{S}_{++}^{n\times n}$, then $\kappa(\boldsymbol{A}) = \lambda_1(\boldsymbol{A})/\lambda_n(\boldsymbol{A})$. Table 1 in [49] compares how the convergence rates of different iterative methods depend on $\kappa(\boldsymbol{A})$. The art of matrix preconditioning is to form an equivalent linear system with an improved condition number which can be solved faster by iterative methods. This is accomplished by multiplying $\boldsymbol{A}$ by simple matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that $\boldsymbol{A}\boldsymbol{X}$, $\boldsymbol{Y}\boldsymbol{A}$ or $\boldsymbol{Y}\boldsymbol{A}\boldsymbol{X}$ have an improved condition number. Preconditioning is a well-studied aspect of numerical linear algebra [9, 57] and numerous approaches have been proposed. For example, a recent paper proposes optimal diagonal preconditioners, i.e., $\boldsymbol{X}$ and $\boldsymbol{Y}$ are diagonal matrices [49]. One popular preconditioning model proposed by Benson [2] is

$$\begin{aligned} &\min \|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{I}\|_F \\ &\text{s.t. } \boldsymbol{X} \in \mathcal{M} \end{aligned} \tag{30}$$

where $\mathcal{M}$ is a special set of matrices. To reduce computation, Benson [2] assumed sufficiently sparse matrices in $\mathcal{M}$, e.g., positive diagonal matrices. We demonstrate the performance of

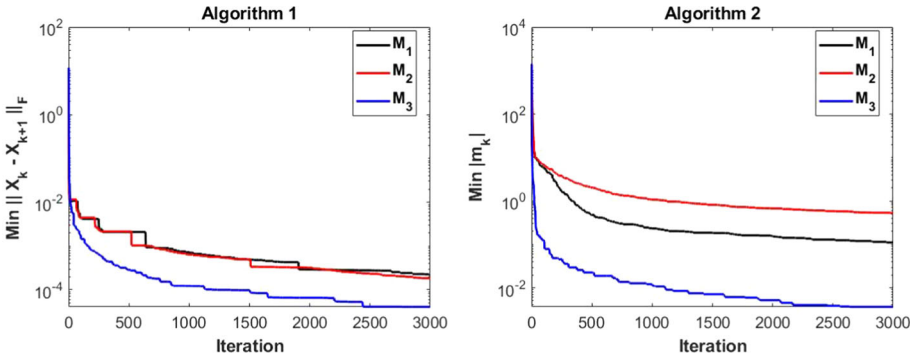

Fig. 2 Convergence plots of the first-order ϵ-stationary condition for Algorithms 1 and 2 applied to (30). The y-axis is in log-scale. For Algorithm 1 the y-axis measures the minimum distance between consecutive iterates at iteration k. For Algorithm 2, the y-axis measures the best solution obtained to (22) at iteration k

our algorithms by solving (30) with various convex eigenvalue constraints, where we do not restrict to sparse matrices:

$$\mathcal{M}_1 := \left\{ \boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \lambda_i(\boldsymbol{X}) \in [0.001, 1],\ i = 1, \ldots, n \right\}, \tag{31}$$

$$\mathcal{M}_2 := \left\{ \boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \lambda_1(\boldsymbol{X}) - \kappa\lambda_n(\boldsymbol{X}) \leq 0,\ \lambda_n(\boldsymbol{X}) \geq 0 \right\}, \tag{32}$$

and

$$\mathcal{M}_3 := \left\{ \boldsymbol{X} \in \mathcal{S}^{n\times n} \mid \boldsymbol{c}_i^\top \lambda(\boldsymbol{X}) \leq 1,\ i = 1, \ldots, m \right\}, \tag{33}$$

where $\kappa > 0$ and $\boldsymbol{c}_i = [i, i-1, \ldots, 1, 0, \ldots, 0]^\top$ for all i. The constraints $\mathcal{M}_1$ and $\mathcal{M}_2$ are reasonable choices for the constraint in (30) because they enforce the preconditioning matrix to be well-conditioned. The first ensures the condition number of the preconditioning matrix $\boldsymbol{X}$ is bounded above by 1000 with bounded eigenvalues between 0.001 and 1; the second ensures the condition number is bounded above by κ while not enforcing an upper bound on the eigenvalues. Note, $\mathcal{M}_2$ does admit the zero matrix as feasible which has an undefined condition number, but this is often of no consequence because often non-zero matrices are present in $\mathcal{M}_2$ which yield better objective function values. The purpose of the last constraint is to showcase the algorithms applied to a general convex set as generated by Theorem 1. The goal of these experiments is to evince the functionality of Algorithms 1 and 2 applied with different constraints on an objective function of practical import. Developing new specialized approaches for preconditioning is outside the scope of this paper.

We applied Algorithms 1 and 2 on three different instances of (30) with $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ serving as the constraint space $\mathcal{M}$. The matrix $\boldsymbol{A} \in \mathcal{S}^{250\times 250}$ to be preconditioned was generated as $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{V}^\top$ where each element of $\boldsymbol{V} \in \mathbb{R}^{250\times 250}$ was drawn from a standard normal. Figure 2 displays the convergence plots of the experiments.

Both methods were initialized at the identity matrix and ran for a total of 3000 iterations. The convergence plots display clearly the sublinear convergence rates guaranteed by Theorems 7 and 9. These numerical results demonstrate the algorithms' performance aligns with the developed theory. We now turn to an application where our methodology outperforms classical approaches.

8.2 Solving Systems of Quadratic Equations

Solving systems of quadratic equations occurs regularly across scientific domains. Two highly studied examples are phase retrieval and the algebraic Riccati equations. Phase retrieval is a problem of recovering a signal from the magnitude of its Fourier transform. Phase retrieval problems are important in imagining science with applications in X-ray crystallography [21, 43], X-ray tomography [12] and astronomy [17], and all phase retrieval problems can be formulated as a system of quadratic equations [5, 6, 27, 56]. The algebraic Riccati equations are crucial in the study of stochastic and optimal control [31], and they also are equivalently expressed as a system of quadratic equations.

Here we demonstrate how to apply our methods to solve systems of quadratic equations. The general form of the problem we consider is:

$$\textbf{Find } \boldsymbol{x} \in \mathbb{R}^n \textbf{ s.t. } \boldsymbol{x}^\top \boldsymbol{Q}_i \boldsymbol{x} = b_i,\ i = 1, \dots m \tag{34}$$

where $\boldsymbol{Q}_i \in \mathbb{R}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. An equivalent matrix form of the problem is:

$$\textbf{Find } \boldsymbol{X} \in \mathcal{S}_+^{n\times n},\ \operatorname{rank}(\boldsymbol{X}) = 1 \textbf{ s.t. } \langle \boldsymbol{Q}_i, \boldsymbol{X}\rangle = b_i,\ i = 1, \dots m$$

which naturally leads to the rank constrained model

$$\begin{aligned} \min\ & \sum_{i=1}^{m} \left(\langle \boldsymbol{Q}_i, \boldsymbol{X}\rangle - b_i\right)^2 \\ \text{s.t. } & \operatorname{rank}(\boldsymbol{X}) = 1 \\ & \boldsymbol{X} \in \mathcal{S}_+^{n\times n}. \end{aligned} \tag{35}$$

If $\boldsymbol{X}^*$ is computed yielding an objective value of zero to (35), then (34) has been solved. If no solution exists to (34), then (35) represents a least-squares approximate solution. A common relaxation of (35) is to drop the rank constraint to obtain a convex model. A global solution to the convex relaxation can be computed and projected onto the set of rank-1 matrices to obtain an approximate solution to (34). Another more precise relaxation of (35) is made possible with our framework. We instead consider the model

$$\begin{aligned} \min\ & \sum_{i=1}^{m} \left(\langle \boldsymbol{Q}_i, \boldsymbol{X}\rangle - b_i\right)^2 \\ \text{s.t. } & \boldsymbol{\lambda}_i(\boldsymbol{X}) \in [0, \delta], \quad i = 2, \dots, n, \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n} \end{aligned} \tag{36}$$

where $\delta > 0$. For small δ, the constraint closely approximates the rank-1 condition. It is easy to check this is a non-convex constraint on the eigenvalues which means the analysis in Sect. 6 does not guarantee convergence to stationary points; however, Theorem 6 guarantees we can compute optimal projections onto the constraint, so we can implement Algorithm 1 on (36).

We considered three methods for solving (34):

Newton: Apply Newton's method directly to (34). In our tests, we initialized and implemented Newton's method ten times and took the best result as the solution.

Convex + Newton: Solve the convex relaxation of (35), project the solution onto the set of rank-1 matrices, and then run Newton's method initialized from the projected solution. In our experiments, we utilized CVX to solve the convex relaxation [18].

SCO + Newton: Apply Algorithm 1 to approximately solve (36), project the solution onto the set of rank-1 matrices, and then run Newton's method initialized from the projected solution.

Our experiments were performed on synthetic data. We simulated random instances of (34) by drawing each entry in the $\boldsymbol{Q}_i$'s from a standard normal distribution and projecting the resulting matrix onto the set of symmetric matrices. The coefficients b_i were determined by randomly selecting a vector $\boldsymbol{y} \in R^n$ with entries from a standard normal and computing $\boldsymbol{y}^\top \boldsymbol{Q}_i \boldsymbol{y}$ for all i. In this way, every system admitted a solution. We conducted ten numerical experiments; five with $(n, m) = (75, 75)$ and five with $(n, m) = (100, 100)$. The maximum allowed iterations for every instance of Newton's method was 5000, and the maximum iterations for Algorithm 1 was 10,000. We used $\delta = 1e-10$ in (36). The error of a potential solution $\boldsymbol{x} \in \mathbb{R}^n$ to (34) was measured as

$$\text{error}(\boldsymbol{x}) = \sum_{i=1}^{m} (\boldsymbol{x}^\top \boldsymbol{Q}_i \boldsymbol{x} - b_i)^2. \tag{37}$$

We initialized Newton's method and Algorithm 1 in two ways. Table 1 displays the results when Newton's method was initialized randomly by sampling from a standard normal distribution and Algorithm 1 was initialized at a fixed diagonal matrix satisfying the constraint in (36). Table 2 provides the results when Newton's method and Algorithm 1 were both initialized near a solution to (34), that is, in all implementations of Newton's method each application of the method was initialized as

$$\boldsymbol{x}^* + \eta \boldsymbol{\sigma}, \tag{38}$$

where $\boldsymbol{x}^*$ was a solution to (34), $\eta > 0$ and each element of $\boldsymbol{\sigma}$ was sampled from a standard normal. Algorithm 1 was initialized at $\boldsymbol{x}_0 \boldsymbol{x}_0^\top$ where $\boldsymbol{x}_0$ was generated by (38). We set $\eta = 0.4$ in our experiments. Note, Convex + Newton does not rely on initialization.

The entries in each table provide the error of the approximated solutions obtained by the three methods above, and the errors obtained by solving the convex relaxation of (35) and solving (36) with Algorithm 1 before application of Newton's method.

In Table 1, we see SCO + Newton significantly outperformed Newton and Convex + Newton in our experiments. The results show SCO + Newton located solutions in three of the ten tests while the other methods failed to compute any solutions to (34). We note also the rank-1 projections of the approximated solutions to (36), obtained by Algorithm 1, given under header "SCO" in Table 1, were always better than the approximated solutions of Newton and Convex + Newton. The vast discrepancy between the projected solutions of the convex relaxation, given under the header "Convex" in Table 1, and (36) demonstrate the over-relaxed nature of removing the rank condition from (35). Though a global minimizer to (36) was not computed, the resulting approximate solution was significantly better than the global minimizer of the convex relaxation, by at least a factor of 10^6. In Table 2, we observe SCO + Newton was able to obtain accurate solutions in nine of the ten experiments when being initialized near a solution to the system of equations. Newton, however, failed to locate any solutions while being initialized in the same neighborhood. We note further in Experiments 4 and 5 when $(n, m) = (75, 75)$ that Algorithm 1 obtained a sufficiently accurate solution to (34) without applying Newton's method. These experiments showcase how our methodology effectively augments the local convergence of Newton's method. Thus, the framework offered by (SCO-Eig) presents a significantly better relaxation of the rank-1 condition which far surpasses the standard convex relaxation. This lends support for our

Table 1 Errors of the three different solvers and the two initial solvers for different synthetic settings

Dimension (n, m)		Newton	Convex	Convex +Newton	SCO	SCO +Newton
(75,75)	Exp. 1	214.21	8.11e8	559.51	1.95	**1.40**
	Exp. 2	331.68	6.30e8	385.27	14.44	**3.80**
	Exp. 3	388.15	7.13e8	471.95	0.03	**7.64e−11**
	Exp. 4	69.09	6.76e8	570.88	37.95	**6.16**
	Exp. 5	154.01	1.03e9	141.76	54.30	**7.37**
(100,100)	Exp. 1	512.87	1.50e9	833.87	1.76	**1.33**
	Exp. 2	929.21	2.18e9	1.22e3	74.86	**8.65**
	Exp. 3	742.94	6.44e9	969.96	0.37	**5.32e−13**
	Exp. 4	660.32	2.05e9	386.52	2.25	**3.06e−10**
	Exp. 5	748.96	4.09e9	832.66	51.74	**7.19**

The column titled "Dimension" states the size of the matrices, $\boldsymbol{Q}_i \in \mathbb{R}^{n\times n}$, and the number of quadratic equations in (34); the rows give the individual experiments for each dimension. The errors (according to (37)) of the proposed methods are given below their respective headers. The column titled "Convex" provides the error of the rank-1 projection of the solution to the convex relaxation of (35); the column "SCO" states the error of the rank-1 projection of the approximate solution to (36) with $\delta = 1e-10$. The bolded numbers mark the best error obtained for each experiment

Table 2 Errors of Newton, SCO + Newton, and one initial solver for different synthetic settings with initialization near an optimal solution

Dimension (n, m)		Newton	SCO	SCO +Newton
(75,75)	Exp. 1	171.20	2.57	**2.62e−9**
	Exp. 2	68.98	0.14	**4.06e−13**
	Exp. 3	200.52	0.01	**7.17e−9**
	Exp. 4	83.27	6.24e−5	**2.40e−10**
	Exp. 5	93.94	1.24e−4	**1.14e−9**
(100,100)	Exp. 1	347.66	0.16	**1.86e−10**
	Exp. 2	207.88	**0.01**	0.09
	Exp. 3	151.35	0.59	**2.18e−13**
	Exp. 4	410.53	0.47	**2.14e−13**
	Exp. 5	193.35	0.12	**2.05e−13**

The column titled "Dimension" states the size of the matrices, $\boldsymbol{Q}_i \in \mathbb{R}^{n\times n}$, and the number of quadratic equations in (34); the rows give the individual experiments for each dimension. The errors (according to (37)) of the proposed methods are given below their respective headers. The column "SCO" states the error of the rank-1 projection of the approximate solution to (36) with $\delta = 1e-10$. The bolded numbers mark the best error obtained for each experiment

framework being a means of relaxing rank constraints, and the results displayed Algorithm 1 can still perform well even when the constraint set is non-convex.

8.3 Matrix Completion

The problem of completing a matrix with missing entries has proved to be useful and important for many tasks including: image reconstruction [60], recommender systems [10], and the

sensor network localization problem [41] among numerous others [16]. Formally, the matrix completion problem can be described as follows: *given a partially observed matrix* $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ *with known entries* $M_{i,j}$ *for* $(i,j) \in \Omega$ *recover the missing entries.* A classical approach to this problem has been to reformulate it as the rank minimization problem

$$\begin{aligned} \min \ & \operatorname{rank}(\boldsymbol{X}) \\ \text{s.t. } & X_{i,j} = M_{i,j}, \ (i,j) \in \Omega \\ & \boldsymbol{X} \in \mathbb{R}^{m\times n}. \end{aligned} \tag{39}$$

This formulation of the problem utilizes the assumption the matrix being recovered has low-rank, which may or may not be the case. A direct approach to solving (39) is well-known to be imprudent because it is in-general NP-hard; so alternative approaches have been proposed. A well-studied approach is to replace the rank function with the nuclear norm and instead solve the convex program

$$\begin{aligned} \min \ & \|\boldsymbol{X}\|_* \\ \text{s.t. } & X_{i,j} = M_{i,j}, \ (i,j) \in \Omega \\ & \boldsymbol{X} \in \mathbb{R}^{m\times n}. \end{aligned} \tag{40}$$

In the seminal work by Candès and Recht [8], they demonstrated solving (40) was able to perfectly recover with high probability low-rank matrices provided enough of the entries were sampled. The main benefit of (40) is the fact it is convex and global minimums of the model are computable; however, as demonstrated in the prior section, sometimes non-convex relaxations yield better solutions than globally solvable convex relaxations. We demonstrate this is also the case for matrix completion when additional information about the spectrum of the desired matrix is known.

Let us consider recovering positive semidefinte matrices. Our set-up is similar to the numerical experiments conducted in [8], but we add the additional piece of information that we know the eigenvalue structure of the matrices being recovered. In detail, given positive integers n and s, with $s \leq n$, we randomly generate $\boldsymbol{M} \in \mathcal{S}_+^{n\times n}$ such that $\lambda_1(\boldsymbol{M}) = \cdots = \lambda_s(\boldsymbol{M}) = n$ and $\lambda_{s+1}(\boldsymbol{M}) = \cdots = \lambda_n(\boldsymbol{M}) = 0$. Leveraging this additional knowledge, we propose the following spectrally constrained matrix completion model

$$\begin{aligned} \min \ & \frac{1}{2}\|\boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{M})\|_F^2 \\ \text{s.t} \quad & \lambda_1(\boldsymbol{X}) = \cdots = \lambda_s(\boldsymbol{X}) = n \\ & \lambda_{s+1}(\boldsymbol{X}) = \cdots = \lambda_n(\boldsymbol{X}) = 0 \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n} \end{aligned} \tag{41}$$

where $W_{ij} = 1$ if $(i,j) \in \Omega$ and $W_{ij} = 0$ otherwise and $\odot$ denotes the Hadamard product. It is clear if a feasible $\boldsymbol{X}$ is computed to (41) such that the objective value is zero a global minimum has been located which satisfies the equality constraints in (40). Also observe, since our theory applies for linear inequality constraints on the spectrum, we can easily deal with equality constraints by adding additional inequality constraints. Thus, (41) is an instance of (SCO-Eig).

In our numerical experiment, we set $n = 50$ and let $s \in \{1, 2, \ldots, 49, 50\}$ and randomly generated data matrices $\boldsymbol{M}$ as described above. We then uniformly at random selected $\alpha \in \{0.05, 0.1, \ldots, 0.90, 0.95\}$ percent of the elements of $\boldsymbol{M}$ to hide. Algorithm 1 was applied to solve (41) and CVX [18] was applied to solve the nuclear norm minimization problem

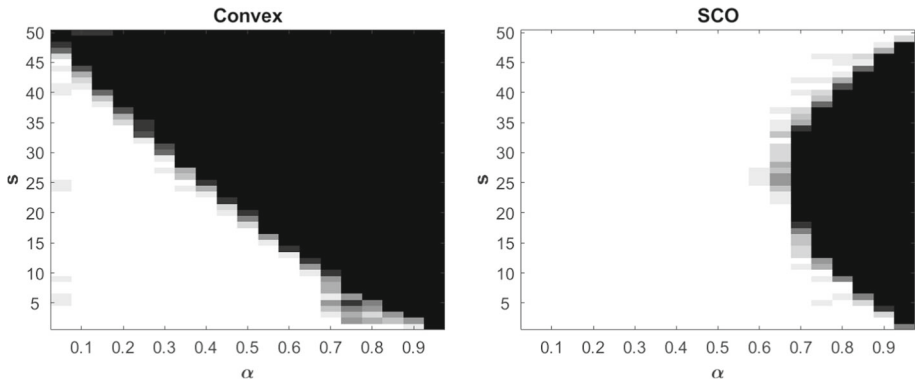

Fig. 3 Displays the results from the matrix recovery experiments. The left panel presents the results from solving the nuclear norm minimization problem (42), while the right panel showcases the results from solving (41); white cells in the figures equate to the method perfectly recovering the matrix in all ten randomized experiments for each setting (s, α) where $s \in \{1, 2, \ldots, 49, 50\}$ and $\alpha \in \{0.05, 0.10, \ldots, 0.90, 0.95\}$; black cells correspond with failure to recover the matrix in any of the ten experiments; gray cells correspond to partial success, i.e., some percentage of the matrices were accurately completed

$$\begin{aligned} \min \ & \mathrm{Tr}(\boldsymbol{X}) \\ \text{s.t. } & X_{i,j} = M_{i,j}, \ (i,j) \in \Omega \\ & \boldsymbol{X} \in \mathcal{S}_+^{n \times n}. \end{aligned} \tag{42}$$

Note, (42) is equivalent to (40) when $\boldsymbol{X}$ is restricted to be positive semidefinite. For all combinations of s and α we conducted ten random experiments. Following [8], we considered a matrix $\boldsymbol{M}$ recovered if the returned matrix $\boldsymbol{X}_{opt}$ satisfied $\|\boldsymbol{X}_{opt} - \boldsymbol{M}\|_F / \|\boldsymbol{M}\|_F < 10^{-3}$. In our tests, we first solved the convex program (42) and projected the solution onto the eigenvalue constraint set in (41) using Theorem 6; from this projected matrix we initialized Algorithm 1 to locate a solution to (41). The results of the experiments are displayed in Fig. 3.

Studying the results, we observe the spectrally constrained approach dominated the classical convex approach by taking into account the eigenvalue structure of the matrices. Of the 9500 numerical tests we conducted represented in the heat maps in Fig. 3 (note white cells equate to perfectly completing the matrix in all ten experiments for the associated (s, α) while black cells equate to failure to recovery the matrix in any of the ten experiments), our procedure recovered the matrices in about 77% of the tests while the convex approach only recovered about 39% of the matrices. Thus, our method nearly doubled the chance of recovery. Though (41) is a highly non-convex problem, our projected gradient method was able to locate global minimums. And by initializing intelligently using the projection of the solution obtained from (42), our method was able to obtain perfect recovery with great frequency.

The regions of recovery approximated in Fig. 3 are also noteworthy. A clear half-circle region in the right panel of Fig. 3 showing failure to recover by solving (41) makes us conjecture a probabilistic recover result is possible for our procedure; this would be an intriguing avenue for future investigation. The key conclusion to draw from this experiment is that when structured eigenvalue information is known non-convex models can significantly best convex relaxations which are incapable of accounting for additional constraints on the spectrum. Also note that though this was a synthetic experiment, examples exist in the

literature where particular structural information about the eigenvalues are known; see for example Theorem 1 of [25]. Such applications are fertile grounds for future investigation.

9 Conclusion

This paper investigates the first matrix optimization model with linear inequality constrained eigenvalues. Theory was developed to understand the nature of the eigenvalue constraint set, and we presented KKT conditions for (SCO-Eig). We additionally verified the accuracy of a relaxation which ensured the differentiability of the eigenvalue operator over the feasible domain. We proved global minima can be computed to (SCO-Eig) for linear objective problems, independent of the convexity of the constraint, and we proved how to compute exact projections. The computational complexity to obtain these global solutions is polynomial and only requires performing a spectral decomposition and solving a simple convex model, e.g., a linear program. Using these results, we developed two algorithms which compute first-order ϵ-stationary points for (SCO-Eig) with a sublinear convergence rate. The practicality of our algorithms was assessed through numerical experimentation. Our example on solving systems of quadratic equations and matrix completion showcased a new and effective method which can best convex relaxations of classical problems.

Many future directions are open for investigation. First, we plan on extending the framework of (SCO-Eig) to include equality constraints on the coordinates of the matrix,

$$\begin{aligned} \min\ & F(\boldsymbol{X}) \\ \text{s.t. } & G_i(\boldsymbol{X}) = 0, i = 1, \ldots, p \\ & \boldsymbol{A}\lambda(\boldsymbol{X}) \leq \boldsymbol{b} \\ & \boldsymbol{X} \in \mathcal{S}^{n\times n}. \end{aligned} \tag{43}$$

This paradigm will encompass most matrix optimization models studied over symmetric matrices and allow new spectral constraints never before considered. Additionally, we plan to develop approaches to solve models which constrain the singular values of non-square matrices and generalized singular values of tensors. As the applications of tensors expand in the burgeoning field of data science, the capability to manipulate the spectrum of these mathematical objects will become ever more important.

Acknowledgements The authors thank the anonymous editor and reviewer for their helpful feedback and comments.

Funding Research Fellowship Program under Grant No. 2237827 and NSF Award DMS-2152766. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Beck, A.: First-Order Methods in Optimization. SIAM (2017)
2. Benson, M.W.: Iterative solution of large sparse linear systems arising in certain multidimensional approximation problems. Util. Math. **22**, 127–140 (1982)
3. Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
4. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
5. Candès, E.J., Eldar, Y.C., Strohmer, T., Voroninski, V.: Phase retrieval via matrix completion. SIAM Rev. **57**(2), 225–251 (2015)
6. Candès, E.J., Li, X.: Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. Found. Comput. Math. **14**, 1017–1026 (2014)
7. Candès, E.J., Plan, Y.: Matrix completion with noise. Proc. IEEE **98**(6), 925–936 (2010)
8. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
9. Chen, K.: Matrix Preconditioning Techniques and Applications. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2005)
10. Chen, Z., Wang, S.: A review on matrix completion for recommender systems. Knowl. Inf. Syst. **64**(1), 1–34 (2022)
11. Cullum, J., Donath, W.E., Wolfe, P.: The Minimization of Certain Nondifferentiable Sums of Eigenvalues of Symmetric Matrices. Nondifferentiable Optimization, pp. 35–55 (1975)
12. Dierolf, M., Menzel, A., Thibault, P., Schneider, P., Kewish, C.M., Wepf, R., Bunk, O., Pfeiffer, F.: Ptychographic x-ray computed tomography at the nanoscale. Nature **467**(7314), 436–439 (2010)
13. Dikin, I.I.: Iterative solution of problems of linear and quadratic programming. In: Doklady Akademii Nauk, Vol. 174, pp. 747–748. Russian Academy of Sciences (1967)
14. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7), 1 (2011)
15. Fan, J., Liao, Y., Liu, H.: An overview of the estimation of large covariance and precision matrices. Economet. J. **19**(1), C1–C32 (2016)
16. Maryam, F.: Matrix Rank Minimization with Applications, Ph.D. thesis,. Stanford University (2002)
17. Fienup, C., Dainty, J.: Phase retrieval and image reconstruction for astronomy. Image recovery: theory and application **231**, 275 (1987)
18. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx (2014)
19. Gowda, M.S.: Optimizing certain combinations of spectral and linear/distance functions over spectral sets. arXiv:1902.06640 (2019)
20. Hager, W.W., Zhang, H.: Projection onto a polyhedron that exploits sparsity. SIAM J. Optim. **26**(3), 1773–1798 (2016)
21. Harrison, R.W.: Phase problem in crystallography. JOSA a **10**(5), 1046–1055 (1993)
22. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**(6), 409–436 (1952)
23. Hiriart-Urruty, J.B., Lewis, A.S.: The Clarke and Michel-Penot subdifferentials of the eigenvalues of a symmetric matrix. Comput. Optim. Appl. **13**, 13–23 (1999)
24. Horn, R.A., Johnson, C.R.: Matrix analysis. Cambridge University Press (1985)
25. Kasten, Y., Geifman, A., Galun, M., Basri, R.: Algebraic characterization of essential matrices and their averaging in multiview settings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5895–5903 (2019)
26. Ito, M., Lourenço, B.F.: Eigenvalue programming beyond matrices. arXiv:2311.04637 (2023)
27. Jaganathan, K., Eldar, Y.C., Hassibi, B.: Phase retrieval: an overview of recent developments. Optical Compressive Imaging, pp. 279–312 (2016)
28. Jourani, A., Ye, J.J.: Error bounds for eigenvalue and semidefinite matrix inequality systems. Math. Program. **104**, 525–540 (2005)
29. Kangal, F., Meerbergen, K., Mengi, E., Michiels, W.: A subspace method for large-scale eigenvalue optimization. SIAM J. Matrix Anal. Appl. **39**(1), 48–82 (2018)
30. Lacoste-Julien, S.: Convergence Rate of Frank-Wolfe for Non-Convex Objectives. arXiv:1607.00345 (2016)
31. Lancaster, P., Rodman, L.: Algebraic Riccati equations. Clarendon Press (1995)
32. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00, pp. 535–541. MIT Press, Cambridge, MA (2000)

33. Lerman, G., Maunu, T.: An overview of robust subspace recovery. Proc. IEEE **106**(8), 1380–1410 (2018)
34. Lewis, A.S.: Derivatives of spectral functions. Math. Oper. Res. **21**(3), 576–588 (1996)
35. Lewis, A.S.: Nonsmooth analysis of eigenvalues. Math. Program. **84**(1), 1–24 (1999)
36. Lewis, A.S.: The mathematics of eigenvalue optimization. Math. Program. **97**, 155–176 (2003)
37. Lewis, A.S., Overton, M.L.: Eigenvalue optimization. Acta numerica **5**, 149–190 (1996)
38. Lewis, A.S., Sendov, H.S.: Twice differentiable spectral functions. SIAM J. Matrix Anal. Appl. **23**(2), 368–386 (2001)
39. Li, Y., Xie, W.: On the Partial Convexification for Low-Rank Spectral Optimization: Rank Bounds and Algorithms. arXiv:2305.07638 (2023)
40. Liang, X., Wang, L., Zhang, L.H., Li, R.C.: On generalizing trace minimization principles. Linear Algebra Appl. **656**, 483–509 (2023)
41. So, A.M.C., Yinyu, Y.: Theory of semidefinite programming for sensor network localization. Math. Program. **109**(2), 367–384 (2007)
42. Mengi, E., Yildirim, E.A., Kilic, M.: Numerical optimization of eigenvalues of Hermitian matrix functions. SIAM J. Matrix Anal. Appl. **35**(2), 699–724 (2014)
43. Millane, R.P.: Phase retrieval in crystallography and optics. JOSA A **7**(3), 394–411 (1990)
44. Ortegar, J.M.: Numerical Analysis. Society for Industrial and Applied Mathematics (1990)
45. Overton, M.L.: Large-scale optimization of eigenvalues. SIAM J. Optim. **2**(1), 88–120 (1992)
46. Overton, M.L., Womersley, R.S.: On minimizing the spectral radius of a nonsymmetric matrix function: Optimality conditions and duality theory. SIAM Matrix Anal. Appt. **9**, 473 (1988)
47. Overton, M.L., Womersley, R.S.: Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. Math. Program. **62**(1–3), 321–357 (1993)
48. Penot, J.-P.: Calculus without derivatives, vol. 266. Springer (2013)
49. Qu, Z., Gao, W., Hinder, O., Ye, Y., Zhou, Z. Optimal Diagonal Preconditioning: Theory and Practice. arXiv:2209.00809 (2022)
50. Saad, Y.: Iterative methods for sparse linear systems. SIAM (2003)
51. Shapiro, A., Fan, M.K.: On eigenvalue optimization. SIAM J. Optim. **5**(3), 552–569 (1995)
52. Tanaka, M., Nakata, K.: Positive definite matrix approximation with condition number constraint. Optim. Lett. **8**(3), 939–947 (2014)
53. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Rev. **38**(1), 49–95 (1996)
54. Vaswani, N., Bouwmans, T., Javed, S., Narayanamurthy, P.: Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. IEEE Signal Process. Mag. **35**(4), 32–55 (2018)
55. Vavasis, S.A., Ye, Y.: A primal-dual interior point method whose running time depends only on the constraint matrix. Math. Program. **74**(1), 79–120 (1996)
56. Wang, G., Giannakis, G.B., Eldar, Y.C.: Solving systems of random quadratic equations via truncated amplitude flow. IEEE Trans. Inf. Theory **64**(2), 773–794 (2017)
57. Wathen, A.J.: Preconditioning. Acta Numer **24**, 329–376 (2015)
58. Won, J.-H., Lim, J., Kim, S.-J., Rajaratnam, B.: Condition-number-regularized covariance estimation. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **75**(3), 427–450 (2013)
59. Ying, Y., Li, P.: Distance Metric Learning with Eigenvalue Optimization. J. Mach. Learn. Res. **13**, 1–26 (2012)
60. Zhang, D., Hu, Y., Ye, J., Li, X., He, X.: Matrix completion by truncated nuclear norm regularization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2192–2199 (2012)
61. Zhang, S.: Global error bounds for convex conic problems. SIAM J. Optim. **10**(3), 836–851 (2000)
62. Zhu, Z., Li, Q., Tang, G., Wakin, M.B.: Global optimality in low-rank matrix optimization. IEEE Trans. Signal Process. **66**(13), 3614–3628 (2018)