GLOBALLY SOLVING THE TRUST REGION SUBPROBLEM USING SIMPLE FIRST-ORDER METHODS*

AMIR BECK† AND YAKOV VAISBOURD‡

Abstract. We consider the trust region subproblem which is given by a minimization of a quadratic, not necessarily convex, function over the Euclidean ball. Based on the well-known second-order necessary and sufficient optimality conditions for this problem, we present two sufficient optimality conditions defined solely in terms of the primal variables. Each of these conditions corresponds to one of two possible scenarios that occur in this problem, commonly referred to in the literature as the presence or absence of the "hard case". We consider a family of first-order methods, which includes the projected and conditional gradient methods. We show that any method belonging to this family produces a sequence which is guaranteed to converge to a stationary point of the trust region subproblem. Based on this result and the established sufficient optimality conditions, we show that convergence to an optimal solution can be also guaranteed as long as the method is properly initialized. In particular, if the method is initialized with the zeros vector and reinitialized with a randomly generated feasible point, then the best of the two obtained vectors is an optimal solution of the problem in probability 1.

Key words. first order methods, trust region subproblem, optimality conditions, global optimum

AMS subject classifications. 90C06, 90C26, 90C46

1. Introduction.

The *trust region subproblem* (TRS) is given by

$$\min\left\{q(\mathbf{x}) : \mathbf{x} \in B\right\}, \tag{TRS}$$

with

$$q(\mathbf{x}) := \mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} \quad \text{and} \quad B := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|^2 \leq 1\},$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric, $\mathbf{b} \in \mathbb{R}^n$ and $\|\cdot\|$ stands for the Euclidean ℓ_2-norm. Though this problem emerges in various applications, it is mostly known due to its central role in the trust region methodology for solving optimization problems, see for example the book [8] and references therein.

In spite of the fact that problem (TRS) is not necessarily convex, its complexity is known to be polynomial [31]. Moreover, problem (TRS) possesses necessary and sufficient second-order optimality conditions which have motivated much of the methods designed to solve it. Probably one of the most popular among those methods is the one suggested in [21], where a safeguarded Newton method is used for finding the root of a secular equation emerging from the aforementioned optimality conditions. This method is applicable for small to medium-sized problems due to the Cholesky factorization which is performed at each iteration.

In the last decades, schemes that maintain modest computational and storage requirements were suggested for addressing large-scale instances of problem (TRS). The conjugate gradient was one of the first iterative methods proposed independently in [26] and [28] for this task. This method succeeds in retrieving the optimal solution

*Submitted to the editors DATE.

Funding: The research of the first author was partially supported by the Israel Science Foundation grant 1821/16.

†School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel (becka@tauex.tau.ac.il).

‡School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel (yakov.vaisbourd@gmail.com).

only if the latter resides in the interior of the feasible set. Otherwise, the scheme breaks down at some stage and the method returns only an approximate solution. In order to continue the process of minimizing over the Krylov subspace from the point in which the conjugate gradient method breaks done, in [14] the authors suggested an adaptation of the Lanczos method that relies on solving at each iteration a much easier TRS, one that involves a tridiagonal matrix in the quadratic term and with a significantly lower dimension. Nevertheless, among other disadvantages of this method, it cannot guarantee to produce the optimal solution either. Trying to overcome these pitfalls, other methods were suggested in the literature such as the sequential subspace method (SSM) [15] which was followed by a variation called phased SSM in [9]. Though these methods can guarantee convergence to the global optimum, to do so they require an estimate of the eigenvector that corresponds to the minimal eigenvalue.

Some schemes for finding the optimal solution of problem (TRS) based on its parameterized eigenvalue reformulation were suggested in the literature. In [25] the authors proposed to examine the close relations between the minimal eigenvector characterization of such a reformulation and the optimality conditions of problem (TRS). Based on this relation, an inverse interpolation scheme for iteratively tuning the parameter is suggested where at each iteration an eigenvalue computation is performed. Various improvements that include a better treatment for the so-called "hard case" (see Subsection 2.3) were suggested in [18, 23]. In [22], a similar parametrized eigenvalue approach was developed using different arguments involving duality theory for SDPs and a homogenization technique.

Under the "difference of convex functions" framework, the projected gradient algorithm was proposed in [27] for solving problem (TRS). In general, there are no guarantees that the sequence generated by the projected gradient method will converge to the global optimal solution, and thus a restarting technique that produces a point with a smaller objective value backed up with a theoretical upper bound on the number of required restarts is suggested in order to avoid convergence to a non-optimal stationary point. Such a restart procedure relies on the availability of an eigenvector that corresponds to the minimal eigenvalue.

Nowadays, TRS for large scale applications is still at the focus of extensive research efforts. Following [11], the work [1] showed how to solve an extension of problem (TRS) using a generalized eigenvector computation. Additional contemporary work include [17, 29] and [16]. The first two independently suggested using an accelerated gradient method for solving an equivalent convex formulation, formed in terms of an eigenvector associated with the minimal eigenvalue, of problem (TRS). In the latter one, problem (TRS) is solved by setting a lower bound on the objective value and adjusting it iteratively. At each iteration a feasibility problem is solved by a semidefinite Programming (SDP) relaxation whose solution involves an eigenvector computation.

In this paper our goal is to show that problem (TRS) can be globally solved by simple first-order methods such as the projected and conditional gradient methods, under proper initialization that can be trivially computed. The dominant computation at each iteration of the class of methods that we consider is multiplication of $\mathbf{A}$ with a given vector, and there is no need to solve or factorize systems of equations or to compute/approximate eigenvectors of $\mathbf{A}$.

The paper's layout is as follows. Section 2 reviews several important mathematical facts that will be pertinent to the analysis throughout the paper: characterizations of stationary and optimality conditions, as well as a discussion on the so-called "easy" and "hard" cases of problem (TRS). Section 3 introduces two sufficient optimality

conditions, one for each case ("easy" and "hard"). These conditions, which are expressed solely in terms of the primal variables, will be instrumental for the proof of convergence to a global optimal solution of the class of "first-order conic methods" (FOCM) introduced in Section 4. This class of algorithms includes the projected and conditional gradient methods as special instances. We show that the entire sequence generated by an FOCM converges regardless of the choice of the starting point. Moreover, it is shown that in the "easy case", an FOCM converges to the optimal solution if initialized with the zeros vector, and in the "hard case", it converges to the global optimal solution as long as the objective function is non-homogenous and the starting point is randomly generated by a continuous distribution whose support is B. A direct consequence is that an optimal solution can be found with probability 1 (even if it is unclear whether the "easy" or "hard" case hold) by employing an FOCM twice–each with a different initialization.

Notation. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{y}$, and matrices by boldface uppercase letters, e.g., $\mathbf{B}$. Given a matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{B}\|_2$ denotes the spectral norm of $\mathbf{B}$, and $\|\mathbf{y}\|$ denotes the ℓ_2 norm of $\mathbf{y}$. The vector $|\mathbf{y}|$ stands for the element-wise absolute value of $\mathbf{y}$. The matrix $\mathbf{I}$ is the identity matrix whose dimension will be clear from the context. The notation $\mathbf{B} \succeq \mathbf{0}$ means that $\mathbf{B}$ is positive semidefinite. The set $\mathbb{N} = \{0, 1, 2 \ldots\}$ is the set of natural numbers.

2. Mathematical Preliminaris.

2.1. Stationarity Conditions.

We begin by recalling (see for example [3, Section 9.1]) that a point $\bar{\mathbf{x}}$ is a *stationary point* of problem (TRS) if it does not possess any feasible descent directions:

$$\nabla q(\bar{\mathbf{x}})^T(\mathbf{x} - \bar{\mathbf{x}}) \geq 0, \text{ for any } \mathbf{x} \in B.$$

It is well-known that the above stationarity condition can be expressed in a more explicit way in the case of problem (TRS).

THEOREM 2.1 ([3, Example 9.6]). *A point $\bar{\mathbf{x}}$ is a stationary point of problem (TRS) if and only if there exists some $\lambda_{(\bar{\mathbf{x}})} \geq 0$ such that the following conditions hold:*

$$(\mathbf{A} + \lambda_{(\bar{\mathbf{x}})}\mathbf{I})\bar{\mathbf{x}} = \mathbf{b}, \tag{2.1}$$

$$(\|\bar{\mathbf{x}}\|^2 - 1)\lambda_{(\bar{\mathbf{x}})} = 0. \tag{2.2}$$

A nonnegative number $\lambda_{(\bar{\mathbf{x}})}$ satisfying (2.1) and (2.2) is called *a Lagrange multiplier associated with* $\bar{\mathbf{x}}$. We also note that condition (2.2) is known as the "complementary slackness" condition. The following simple lemma shows that each stationary point has a *unique* Lagrange multiplier. This uniqueness result on the associated Lagrange multiplier shows that the notation "$\lambda_{(\bar{\mathbf{x}})}$", that will be used throughout the paper, is well-defined.

LEMMA 2.2. *Any stationary point $\bar{\mathbf{x}}$ of problem (TRS) admits a unique Lagrange multiplier $\lambda_{(\bar{\mathbf{x}})} \geq 0$.*

Proof. Assume by contradiction that $\bar{\mathbf{x}}$ admits two Lagrange multipliers $\lambda^1_{(\bar{\mathbf{x}})}$ and $\lambda^2_{(\bar{\mathbf{x}})}$ such that $\lambda^1_{(\bar{\mathbf{x}})} \neq \lambda^2_{(\bar{\mathbf{x}})}$. Then by (2.1)

$$\lambda^1_{(\bar{\mathbf{x}})}\bar{\mathbf{x}} = \mathbf{b} - \mathbf{A}\bar{\mathbf{x}} = \lambda^2_{(\bar{\mathbf{x}})}\bar{\mathbf{x}}.$$

Since $\lambda^1_{(\bar{\mathbf{x}})} \neq \lambda^2_{(\bar{\mathbf{x}})}$ we obtain that $\bar{\mathbf{x}} = \mathbf{0}$, but in this case, condition (2.2) implies that $\lambda^1_{(\bar{\mathbf{x}})} = \lambda^2_{(\bar{\mathbf{x}})} = 0$, leading to a contradiction. ∎

It is interesting to note that the converse of Lemma 2.2 does not hold true in general, meaning that it might happen that the same Lagrange multiplier will be associated with different stationary points. For example, if $q(\mathbf{x}) := 0$ ($\mathbf{A} = \mathbf{0}, \mathbf{b} = \mathbf{0}$), then all the feasible points are stationary points, and $\lambda = 0$ is the Lagrange multiplier associated will all of them.

Obviously, stationarity is a necessary optimality condition [6]. Combining this fact with Lemma 2.2 yields the following known result.

THEOREM 2.3. *Any optimal solution $\bar{\mathbf{x}}$ of problem (TRS) is also a stationary point, meaning that $(\bar{\mathbf{x}}, \lambda_{(\bar{\mathbf{x}})})$ satisfies conditions (2.1) and (2.2).*

Problem (TRS) admits necessary and sufficient second-order optimality conditions which are summarized in the following theorem (see e.g., [12, 24]). The conditions state that a point is optimal if and only if it is stationary and satisfies a certain second-order optimality condition.

THEOREM 2.4 ([24, Lemmas 2.4 and 2.8]). *Let $\bar{\mathbf{x}}$ be a feasible point of problem (TRS). Then $\bar{\mathbf{x}}$ is an optimal solution of problem (TRS) if and only if it is a stationary point such that*

$$\mathbf{A} + \lambda_{(\bar{\mathbf{x}})}\mathbf{I} \succeq \mathbf{0}. \tag{2.3}$$

2.2. Spectral Characterization of Optimality and Stationarity. Throughout the analysis in the paper we will consider the stationarity and optimality conditions in terms of the eigenvalues and eigenvectors of $\mathbf{A}$. Let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be a spectral decomposition of $\mathbf{A}$ such that $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ with $d_1 \leq d_2 \leq \cdots \leq d_n$ being the increasingly sorted eigenvalues of $\mathbf{A}$ and $\mathbf{U}$ being an orthonormal matrix whose columns are associated eigenvectors of $\mathbf{A}$. **The notations of the matrix $\mathbf{U}$ and $d_1 \leq d_2 \leq \cdots \leq d_n$ will be fixed throughout the paper.**
In terms of the spectral decomposition, Theorems 2.1 and 2.4 as well as Lemma 2.2 can be rewritten in the following compact form.

THEOREM 2.5. (a) *A point $\bar{\mathbf{x}}$ is a stationary point of problem (TRS) if and only if there exists a unique $\lambda_{(\bar{\mathbf{x}})} \geq 0$ such that the following conditions hold:*

$$(d_i + \lambda_{(\bar{\mathbf{x}})})(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\mathbf{b})_i, \quad i = 1, 2, \ldots, n. \tag{2.4}$$

$$(\|\bar{\mathbf{x}}\|^2 - 1)\lambda_{(\bar{\mathbf{x}})} = 0. \tag{2.5}$$

(b) *A feasible point $\bar{\mathbf{x}}$ of problem (TRS) is optimal if and only if it is a stationary point and satisfies*

$$\lambda_{(\bar{\mathbf{x}})} \geq -d_1.$$

2.3. The "Hard Case". The characterization of an optimal solution given by the three conditions (2.1)–(2.3) motivated the development of some efficient methods for solving problem (TRS), see for example [21, 24] and the discussion in the introduction. These methods, as well as many algorithms that appear in the literature, provide a special treatment for the so-called "hard case" which we now describe. Denote the eigenspace of $\mathbf{A}$ corresponding to the smallest eigenvalue, which we denote as d_1, by

$$\mathcal{E}_1 := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = d_1\mathbf{x}\}.$$

There are two classes of trust region subproblems that are frequently analyzed separately. The classes are determined according to whether or not the vector $\mathbf{b}$ is orthogonal to the subspace $\mathcal{E}_1$

- **"easy case"** $\mathbf{b} \not\perp \mathcal{E}_1$.
- **"hard case"** $\mathbf{b} \perp \mathcal{E}_1$.

We will often use the following notation for the set of indices of eigenvalues corresponding to the minimal eigenvalue:

$$E_1 := \{i \in \{1, 2, \ldots, n\} : d_i = d_1\}. \tag{2.6}$$

With the above notation, the condition describing the "easy case" can also be written as

$$(\mathbf{U}^T\mathbf{b})_i \neq 0 \text{ for some } i \in E_1.$$

It is well known that in the "easy case" there exists a unique optimal solution to problem (TRS).
The "hard case" is, in a sense, not likely to occur since, loosely speaking, for randomly generated data ($\mathbf{A}$ and $\mathbf{b}$), the probability that the "hard case" will occur is 0. Therefore, the condition $\mathbf{b} \not\perp \mathcal{E}_1$ describing the "easy case" is considered to be mild. However, theoretically, the "hard case" might arise, and it usually requires a more involved and complicated analysis than the one required for the "easy case" (hence, the name).

3. Two Sufficient Optimality Conditions. In this section we will present two sufficient optimality conditions that will be the basis for showing how a global optimal solution of problem (TRS) can be obtained in both the "hard" and "easy" cases.
The sufficient conditions require that a stationary point will reside in some set. In the "easy case" the relevant set is

$$S_E = \{\mathbf{x} \in B : \operatorname{sign}((\mathbf{U}^T\mathbf{b})_i)(\mathbf{U}^T\mathbf{x})_i \geq 0, i \in I_-\}, \tag{3.1}$$

where $\operatorname{sgn}(\alpha)$ denotes the sign function which returns 1 if $\alpha \geq 0$ and -1 otherwise and

$$I_- = \{i \in \{1, 2, \ldots, n\} : d_i \leq 0\}.$$

In the "hard case", the relevant set is

$$S_H := \{\mathbf{x} \in B : \mathbf{x} \not\perp \mathcal{E}_1\}. \tag{3.2}$$

We are now ready to prove a sufficient optimality condition for each of the scenarios.

THEOREM 3.1 (**sufficient optimality condition - "easy case"**). *Suppose that the "easy case" holds. Let $\bar{\mathbf{x}} \in S_E$ be a stationary point of problem (TRS). Then $\bar{\mathbf{x}}$ is the optimal solution of problem (TRS).*

Proof. If $d_1 \geq 0$, then problem (TRS) is convex, and obviously any stationary point is optimal. Therefore, we will henceforth assume that $d_1 < 0$. Since $\bar{\mathbf{x}}$ is a stationary point of problem (TRS), then by Theorem 2.5(a) for any $i = 1, 2, \ldots, n$,

$$(d_i + \lambda_{(\bar{\mathbf{x}})})(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\mathbf{b})_i. \tag{3.3}$$

Since the "easy case" holds, it follows that there exists $i_1 \in E_1$ such that $(\mathbf{U}^T\mathbf{b})_{i_1} \neq 0$, which by (3.3) implies that $(\mathbf{U}^T\bar{\mathbf{x}})_{i_1} \neq 0$. Combining this with the fact that $\bar{\mathbf{x}} \in S_E$

and $d_1 < 0$, we obtain that $(\mathbf{U}^T\bar{\mathbf{x}})_{i_1}(\mathbf{U}^T\mathbf{b})_{i_1} > 0$. Therefore, by (3.3) with $i = i_1$,

$$\lambda_{(\bar{\mathbf{x}})} = -d_{i_1} + \frac{(\mathbf{U}^T\mathbf{b})_{i_1}}{(\mathbf{U}^T\bar{\mathbf{x}})_{i_1}} > -d_{i_1} = d_1,$$

which in light of Theorem 2.5(b) implies that $\bar{\mathbf{x}}$ is an optimal solution of problem (TRS). ∎

Remark 3.2. The condition defining the set S_E, namely "$\operatorname{sgn}((\mathbf{U}^T\mathbf{b})_i)(\mathbf{U}^T\bar{\mathbf{x}})_i \geq 0$" corresponds to a property that is known to hold for optimal solutions of problem (TRS), and is the basis for its so-called "hidden convexity" property, see [5] as well as [3, Lemma 8.7].

THEOREM 3.3 (**sufficient optimality condition - "hard case").** *Suppose that the "hard case" holds. Let $\bar{\mathbf{x}} \in S_H$ be a stationary point of problem (TRS). Then $\bar{\mathbf{x}}$ is an optimal solution of problem (TRS).*

Proof. Since $\bar{\mathbf{x}} \in S_H$, it holds that $\bar{\mathbf{x}} \notin \mathcal{E}_1$, i.e., there exists $i \in E_1$ such that $(\mathbf{U}^T\bar{\mathbf{x}})_i \neq 0$. For such an i, the "hard case" assumption implies that $(\mathbf{U}^T\mathbf{b})_i = 0$. By Theorem 2.5(a), we obtain that

$$(\lambda_{(\bar{\mathbf{x}})} + d_i)(\mathbf{U}^T\bar{\mathbf{x}})_i = 0,$$

which combined with the fact that $(\mathbf{U}^T\bar{\mathbf{x}})_i \neq 0$ implies that $\lambda_{(\bar{\mathbf{x}})} = -d_i = -d_1$, and hence, by Theorem 2.5(b), $\bar{\mathbf{x}}$ is an optimal solution of problem (TRS). ∎

In the next section we will show how the two sufficient conditions described in Theorems 3.1 and 3.3 form the basis for the convergence to a global optimal solution of a class of simple first-order methods that includes, among others, the projected gradient and conditional gradient methods. Specifically, it will be shown that in the "easy case", an algorithm that belongs to this class of methods, initialized at a point in S_E, converges to a stationary point in S_E, which by Theorem 3.1, is also optimal. In the "hard case", based on Theorem 3.3, it will be shown that convergence to an optimal solution is guaranteed as long as the initial point belongs to S_H. If it is not clear (as is the common situation) whether the hard or easy cases hold, then a global optimal solution can be obtained by running the method twice–once with $\mathbf{x}^0 \in S_E$ and the second time with $\mathbf{x}^0 \in S_H$. Note that although both S_E and S_H are described in terms of the spectral decomposition of $\mathbf{A}$, an operation that should be avoided, we will show that there is a simple way to choose the initial vector $\mathbf{x}^0$ without the need to access any spectral information.

4. First-Order Conic Methods.

4.1. Definition and Examples.

We begin with the definition of a "first-order conic method".

DEFINITION 4.1 (**first-order conic method).** *A method which generates a sequence of points $\{\mathbf{x}^k\}_{k=0}^{\infty}$ according to the general step*

$$\mathbf{x}^{k+1} = \theta_1^k\mathbf{x}^k + \theta_2^k\left(-\nabla q(\mathbf{x}^k)\right). \tag{4.1}$$

with some $\mathbf{x}^0 \in B$ is called a first-order conic method *(FOCM) if it satisfies the following conditions:*

(A) The method generates a feasible sequence, i.e., $\mathbf{x}^k \in B$ for any $k \in \mathbb{N}$.

(B) $q(\mathbf{x}^{k+1}) \leq q(\mathbf{x}^k)$ for all $k \in \mathbb{N}$.

(C) All limit points of the sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ *are stationary points of problem* (TRS).
(D) $\boldsymbol{\theta}^k \in \mathbb{R}^2_+ \setminus \{\mathbf{0}\}$ *and* $\theta_1^k \in [0,1]$ *for any* $k \in \mathbb{N}$.

Property (D) in the definition of an FOCM above implies that

$$\mathbf{x}^{k+1} \in B \cap \operatorname{cone}\left(\{\mathbf{x}^k, -\nabla q(\mathbf{x}^k)\}\right),$$

which explains the name of the class of methods. Note that Definition 4.1 is quite general. Following are two examples of algorithms which are FOCMs.

4.1.1. Projected Gradient. The projected gradient (PG) method is one of the most fundamental methods for solving differentiable constrained optimization problems [6, 19]. In the case of problem (TRS), the general step of the PG method is given by

Projected Gradient (PG) Method

For any $k = 0, 1, 2, \ldots$, execute the following:

$$\mathbf{x}^{k+1} := P_B(\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)) = \begin{cases} \mathbf{x}^k - t_k \nabla q(\mathbf{x}^k), & \|\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)\| \leq 1, \\ \frac{\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)}{\|\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)\|}, & \text{otherwise}, \end{cases}$$

where $t_k \geq 0$ is a corresponding step-size and

$$P_B(\mathbf{y}) := \operatorname{argmin}\{\|\mathbf{x} - \mathbf{y}\|^2 : \mathbf{x} \in B\} = \begin{cases} \mathbf{y}, & \|\mathbf{y}\| \leq 1, \\ \frac{\mathbf{y}}{\|\mathbf{y}\|}, & \|\mathbf{y}\| > 1. \end{cases}$$

is the orthogonal projection operator. This method, with a constant step size,was previously suggested for solving problem (TRS) in [27] where a restarting procedure, that involves an eigenvector computation, was used in order to guarantee convergence to the optimal solution.

Property (A) in the definition of an FOCM (feasibility) is surely satisfied because of the projection step. Property (D) also holds since $t_k \geq 0$ and $\theta_1^k \in [0,1]$ since it is either equal to 1 if $\|\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)\| \leq 1$ and if $\|\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)\| > 1$, then $\theta_1^k = \|\mathbf{x}^k - t_k \nabla q(\mathbf{x}^k)\|^{-1} < 1$. The validity of properties (B) and (C) can be shown under some step-size regimes. Here we consider the following two popular step-size strategies:

- **Constant.** $t_k = \frac{1}{\bar{L}}$, where $\bar{L} \in [\|\mathbf{A}\|_2, \infty)$.
- **Backtracking.** The procedure requires three parameters (s, γ, η) where $s > 0$, $\gamma \in (0,1)$ and $\eta > 1$. The step-size is chosen as $t_k = \frac{1}{L_k}$ where L_k is picked as follows: first, L_k is set to be equal to the initial guess s. Then, while

$$q(\mathbf{x}^k) - q\left(P_B\left(\mathbf{x}^k - \frac{1}{L_k}\nabla q(\mathbf{x}^k)\right)\right) < \gamma L_k \left\|\mathbf{x}^k - P_B\left(\mathbf{x}^k - \frac{1}{L_k}\nabla q(\mathbf{x}^k)\right)\right\|^2,$$

we set $L_k := \eta L_k$.

It is well known (see for example [4, Theorem 10.15]) that properties (B) and (C) hold for the PG method under the constant and backtracking strategies as described above. We can thus conclude the following.

THEOREM 4.2. *The PG method with either constant or backtracking step-size strategies is an FOCM.*

4.1.2. Conditional Gradient. Another fundamental method for solving differentiable and constrained optimization problems is the conditional gradient (CG) method [6, 10, 19]. In the case of problem (TRS), the CG method takes the following form.

Conditional Gradient (CG) Method
For any $k = 0, 1, 2, \ldots$, execute the following:
(a) compute $\mathbf{p}^k \in \underset{\mathbf{p}\in B}{\operatorname{argmin}} \langle \mathbf{p}, \nabla q(\mathbf{x}^k)\rangle$;
(b) choose $t_k \in [0, 1]$ and set $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}^k - \mathbf{x}^k)$.

Step (a) of the method is not well-defined when $\nabla q(\mathbf{x}^k) = \mathbf{0}$ since in this case we can choose $\mathbf{p}^k$ to be any point in B. To make the method well-defined, and in order to avoid unnecessary pathological situations, we will set $\mathbf{p}^k = \mathbf{0}$ and $t_k = 0$ in this case. With this convention, we have

$$\mathbf{p}^k = \begin{cases} -\frac{\nabla q(\mathbf{x}^k)}{\|\nabla q(\mathbf{x}^k)\|}, & \nabla q(\mathbf{x}^k) \neq \mathbf{0}, \\ \mathbf{0}, & \nabla q(\mathbf{x}^k) = \mathbf{0}. \end{cases}$$

and thus, the general update step is given by

$$\mathbf{x}^{k+1} = \begin{cases} (1 - t_k)\mathbf{x}^k - \frac{t_k}{\|\nabla q(\mathbf{x}^k)\|}\nabla q(\mathbf{x}^k), & \nabla q(\mathbf{x}^k) \neq \mathbf{0}, \\ \mathbf{x}^k, & \nabla q(\mathbf{x}^k) = \mathbf{0}. \end{cases}$$

The above representation shows that the CG method satisfies properties (A) and (D) required from an FOCM. We will consider the following step-size strategy:

- **Exact line search.** $t_k \in \underset{t\in[0,1]}{\operatorname{argmin}}\, q(\mathbf{x}^k + t(\mathbf{p}^k - \mathbf{x}^k))$.

The CG method with exact line search strategy is known to satisfy properties (B) and (C) in the definition of an FOCM (see for example [4, Chapter 13]), and we can thus conclude that in this setting, the CG method is an FOCM.

THEOREM 4.3. *The CG method with an exact line-search strategy is an FOCM.*

Remark 4.4. The constant step-size selection rule suggested for the GP method, involve an upper bound on the spectral norm of $\mathbf{A}$. In some cases such an upper bound can be easily and efficiently acquired. However, in general, since $\|\mathbf{A}\|_2 = \max\{|d_1|, |d_n|\}$, assessing the spectral norm requires an eigenvalue computation. This, of course, can be avoided by using the backtracking step-size selection rule, which do not require any spectral information.

4.2. Convergence of the Sequence Generated by an FOCM. Our main objective will be to show how to achieve the global optimal point of problem (TRS) using any FOCM method. Before that, we will show that the sequence generated by an FOCM converges to a stationary point. This is a stronger property than condition (C) in the definition of an FOCM (Definition 4.1), which only warrants that limit points of the generated sequence are stationary point. This result will also be useful in showing later on how convergence to an optimal solution can be achieved under proper initialization.

THEOREM 4.5. *Let $\{\mathbf{x}^k\}_{k=0}^{\infty}$ be a sequence generated by an FOCM. Then $\{\mathbf{x}^k\}_{k=0}^{\infty}$ converges to a stationary point of problem (TRS).*

Proof. Due to the monotonicity of $\{q(\mathbf{x}^k)\}_{k=0}^{\infty}$ (property (B) in Definition 4.1) and the fact that problem (TRS) admits an optimal solution, we can conclude that there is some $\bar{q}$ such that

$$\lim_{k\to\infty} q(\mathbf{x}^k) = \bar{q}. \tag{4.2}$$

Since by the definition of an FOCM (specifically, property (C) in Definition 4.1), any limit point of the sequence is a stationary point, all we need to prove is that the sequence converges. Assume by contradiction that the sequence does not converge. Let then $\bar{\mathbf{x}}, \bar{\mathbf{y}} \in B$, $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ be two different accumulation points of the sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$. By (4.2) and the continuity of q, we obtain that $q(\bar{\mathbf{x}}) = q(\bar{\mathbf{y}}) = \bar{q}$. In addition, by property (C) in the definition of an FOCM (Definition 4.1), both points are stationary points of problem (TRS). We will show that $\lambda_{(\bar{\mathbf{x}})} = \lambda_{(\bar{\mathbf{y}})}$. Indeed,

$$\begin{aligned} q(\bar{\mathbf{x}}) - q(\bar{\mathbf{y}}) &= \bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}} - 2\mathbf{b}^T\bar{\mathbf{x}} - \bar{\mathbf{y}}^T\mathbf{A}\bar{\mathbf{y}} + 2\mathbf{b}^T\bar{\mathbf{y}} \\ &= (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{A}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) + 2(\mathbf{A}\bar{\mathbf{y}} - \mathbf{b})^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) \\ &= (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{A}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - 2\lambda_{(\bar{\mathbf{y}})}\bar{\mathbf{y}}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}), \end{aligned}$$

where in the last equality we have used (2.1) due to the stationarity of $\bar{\mathbf{y}}$. Similarly, we obtain that

$$q(\bar{\mathbf{y}}) - q(\bar{\mathbf{x}}) = (\bar{\mathbf{y}} - \bar{\mathbf{x}})^T\mathbf{A}(\bar{\mathbf{y}} - \bar{\mathbf{x}}) - 2\lambda_{(\bar{\mathbf{x}})}\bar{\mathbf{x}}^T(\bar{\mathbf{y}} - \bar{\mathbf{x}}).$$

Since we already established that $q(\bar{\mathbf{x}}) = q(\bar{\mathbf{y}})$, by combining the last two results we obtain that

$$\lambda_{(\bar{\mathbf{x}})}(\bar{\mathbf{y}}^T\bar{\mathbf{x}} - \|\bar{\mathbf{x}}\|^2) = \lambda_{(\bar{\mathbf{y}})}(\bar{\mathbf{y}}^T\bar{\mathbf{x}} - \|\bar{\mathbf{y}}\|^2). \tag{4.3}$$

If $\|\bar{\mathbf{x}}\|^2 = \|\bar{\mathbf{y}}\|^2$ then due to our assumption that $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$, we obtain that[1] $\bar{\mathbf{y}}^T\bar{\mathbf{x}} < \|\bar{\mathbf{x}}\|^2 = \|\bar{\mathbf{y}}\|^2$ and thus $\lambda_{(\bar{\mathbf{x}})} = \lambda_{(\bar{\mathbf{y}})}$. Otherwise, assume without loss of generality that $\|\bar{\mathbf{x}}\|^2 < \|\bar{\mathbf{y}}\|^2$ which implies by the complementary slackness condition (2.5) that $\lambda_{(\bar{\mathbf{x}})} = 0$. Then (4.3) boils down to $\lambda_{(\bar{\mathbf{y}})}(\bar{\mathbf{y}}^T\bar{\mathbf{x}} - \|\bar{\mathbf{y}}\|^2) = 0$ and since $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ we obtain that[2] $\lambda_{(\bar{\mathbf{y}})} = 0$.

To summarize, up to this point we assumed by contradiction that the sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ has two different accumulation points $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ and showed that $\lambda_{(\bar{\mathbf{x}})} = \lambda_{(\bar{\mathbf{y}})}$. We will denote the common Lagrange multiplier value by $\lambda := \lambda_{(\bar{\mathbf{x}})} = \lambda_{(\bar{\mathbf{y}})}$. Now, since both $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are stationary points of problem (TRS), then by Theorem 2.5(a)

$$(\lambda + d_i)(\mathbf{U}^T\mathbf{z})_i = (\mathbf{U}^T\mathbf{b})_i, \qquad \mathbf{z} \in \{\bar{\mathbf{x}}, \bar{\mathbf{y}}\},\ i = 1, 2, \ldots, n. \tag{4.4}$$

If $\lambda \neq -d_i$ for all $i = 1, 2, \ldots, n$, then by (4.4)

$$(\mathbf{U}^T\bar{\mathbf{x}})_i = \frac{(\mathbf{U}^T\mathbf{b})_i}{\lambda + d_i} = (\mathbf{U}^T\bar{\mathbf{y}})_i, \quad i = 1, 2, \ldots, n,$$

[1] Since $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ and $\|\bar{\mathbf{x}}\| = \|\bar{\mathbf{y}}\|$, it follows that $\bar{\mathbf{x}}, \bar{\mathbf{y}} \neq \mathbf{0}$. To prove the strict inequality, note that by the Cauchy-Schwarz inequality $\bar{\mathbf{y}}^T\bar{\mathbf{x}} \leq \|\bar{\mathbf{y}}\| \cdot \|\bar{\mathbf{x}}\| = \|\bar{\mathbf{x}}\|^2$. Equality will hold only if $\bar{\mathbf{y}} = \alpha\bar{\mathbf{x}}$ for some nonnegative α, but since $\|\bar{\mathbf{x}}\| = \|\bar{\mathbf{y}}\|$, α has to be equal to 1, leading to a contradiction to the assumption that $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$.

[2] Indeed, by the Cauchy-Schwarz inequality $\bar{\mathbf{y}}^T\bar{\mathbf{x}} \leq \|\bar{\mathbf{y}}\| \cdot \|\bar{\mathbf{x}}\| < \|\bar{\mathbf{y}}\|^2$, where the strict inequality is due to the facts that $\bar{\mathbf{y}} \neq \mathbf{0}$ and $\|\bar{\mathbf{x}}\| < \|\bar{\mathbf{y}}\|$.

implying that $\bar{\mathbf{x}} = \bar{\mathbf{y}}$, which is a contradiction to the assumption that $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$. We will therefore assume from now on that there is some $m \in \{1, 2, \ldots, n\}$ for which $\lambda_{(\bar{\mathbf{x}})} = -d_m$. We denote (expanding the notation "E_1", see (2.6)),

$$E_m = \{\ell \in \{1, 2, \ldots, n\} : d_\ell = d_m\}.$$

For any $i \notin E_m$, by (4.4),

$$(\mathbf{U}^T\bar{\mathbf{x}})_i = \frac{(\mathbf{U}^T\mathbf{b})_i}{\lambda + d_i} = (\mathbf{U}^T\bar{\mathbf{y}})_i.$$

We will show that $(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\bar{\mathbf{y}})_i$ also for any $i \in E_m$. Let then $i \in E_m$. Then due to (4.4), since $\lambda = -d_m$, it must hold that $(\mathbf{U}^T\mathbf{b})_i = 0$, and thus, the update step (4.1) of the FOCM method is (recalling that $d_i = d_m$ for $i \in E_m$),

$$\begin{aligned}(\mathbf{U}^T\mathbf{x}^{k+1})_i &= \theta_1^k(\mathbf{U}^T\mathbf{x}^k)_i + 2\theta_2^k\left(-d_i(\mathbf{U}^T\mathbf{x}^k)_i\right)\\ &= \left[\prod_{j=0}^k(\theta_1^j - 2\theta_2^j d_m)\right](\mathbf{U}^T\mathbf{x}^0)_i,\end{aligned}$$

meaning that

$$(\mathbf{U}^T\mathbf{x}^{k+1})_i = \pi_k(\mathbf{U}^T\mathbf{x}^0)_i, \tag{4.5}$$

where

$$\pi_k := \prod_{j=0}^k(\theta_1^j - 2\theta_2^j d_m).$$

Note that $\pi_k \geq 0$ for any k since $-d_m = \lambda \geq 0$ and $\theta_1^j, \theta_2^j \geq 0$.

We will now consider two cases.

(a) If $\lambda = 0$, then for any $i \in E_m$ (4.5) is simply

$$(\mathbf{U}^T\mathbf{x}^{k+1})_i = \pi_k(\mathbf{U}^T\mathbf{x}^0)_i,$$

where here $\pi_k = \prod_{j=0}^k \theta_1^j$. Since $\theta_1^k \in [0, 1]$ for any $k \in \mathbb{N}$, the sequence $\{\pi_k\}_{k=0}^\infty$ is non-increasing and bounded from below by zero, and thus converges to some β. Hence, $\{(\mathbf{U}^T\mathbf{x}^k)_i\}_{k=0}^\infty$ converges to $\beta(\mathbf{U}^T\mathbf{x}^0)_i$. Since $(\mathbf{U}^T\bar{\mathbf{x}})_i$ and $(\mathbf{U}^T\bar{\mathbf{y}})_i$ are limit points of the convergent sequence $\{(\mathbf{U}^T\mathbf{x}^k)_i\}_{k=0}^\infty$, it follows that $(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\bar{\mathbf{y}})_i$.

(b) If $\lambda \neq 0$, then due to the complementary slackness condition and the fact, previously established, that $(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\bar{\mathbf{y}})_i$ for any $i \notin E_m$, we obtain that

$$\sum_{i\in E_m}(\mathbf{U}^T\bar{\mathbf{x}})_i^2 = \sum_{i\in E_m}(\mathbf{U}^T\bar{\mathbf{y}})_i^2. \tag{4.6}$$

By (4.5) it follows that $(\mathbf{U}^T\bar{\mathbf{x}})_i(\mathbf{U}^T\bar{\mathbf{y}})_i \geq 0$ for any $i \in E_m$ and consequently, the fact that $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ implies that $|\mathbf{U}^T\bar{\mathbf{x}}| \neq |\mathbf{U}^T\bar{\mathbf{y}}|$. Combining this with (4.6), it follows that there must be some $i_1, i_2 \in E_m$ and some $\epsilon > 0$ such that

$$|(\mathbf{U}^T\bar{\mathbf{x}})_{i_1}| - |(\mathbf{U}^T\bar{\mathbf{y}})_{i_1}| > 2\epsilon \quad \text{and} \quad |(\mathbf{U}^T\bar{\mathbf{y}})_{i_2}| - |(\mathbf{U}^T\bar{\mathbf{x}})_{i_2}| > 2\epsilon. \tag{4.7}$$

Since $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are accumulation points of the sequence $\{\mathbf{x}^k\}_{k=0}^\infty$, there are two subsequences $\{\mathbf{x}^{k_t^1}\}_{t=0}^\infty$ and $\{\mathbf{x}^{k_t^2}\}_{t=0}^\infty$ such that $\mathbf{x}^{k_t^1} \to \bar{\mathbf{x}}$ and $\mathbf{x}^{k_t^2} \to \bar{\mathbf{y}}$ as $t \to \infty$. Hence there exists $t_1 \in \mathbb{N}$ such that

$$|(\mathbf{U}^T\mathbf{x}^{k_t^1})_{i_1} - (\mathbf{U}^T\bar{\mathbf{x}})_{i_1}| < \epsilon \text{ and } |(\mathbf{U}^T\mathbf{x}^{k_t^1})_{i_2} - (\mathbf{U}^T\bar{\mathbf{x}})_{i_2}| < \epsilon \text{ for any } t \geq t_1. \tag{4.8}$$

Similarly, there exists $t_2 \in \mathbb{N}$ such that

(4.9)
$$|(\mathbf{U}^T\mathbf{x}^{k_t^2})_{i_1} - (\mathbf{U}^T\bar{\mathbf{y}})_{i_1}| < \epsilon \text{ and } |(\mathbf{U}^T\mathbf{x}^{k_t^2})_{i_2} - (\mathbf{U}^T\bar{\mathbf{y}})_{i_2}| < \epsilon \text{ for any } t \geq t_2.$$

Assume, without the loss of generality, that $k_{t_2}^2 > k_{t_1}^1$. Combining (4.7) and (4.8), we obtain

$$|(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_1}| - |(\mathbf{U}^T\bar{\mathbf{y}})_{i_1}| > \epsilon, \tag{4.10}$$

$$|(\mathbf{U}^T\bar{\mathbf{y}})_{i_2}| - |(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_2}| > \epsilon. \tag{4.11}$$

Due to (4.5) we also have that

$$(\mathbf{U}^T\mathbf{x}^{k_{t_2}^2})_{i_1} = \eta(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_1} \quad \text{and} \quad (\mathbf{U}^T\mathbf{x}^{k_{t_2}^2})_{i_2} = \eta(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_2}, \tag{4.12}$$

where

$$\eta = \left[\prod_{j=k_{t_1}^1}^{k_{t_2}^2-1} (\theta_1^j - 2\theta_2^j d_m)\right].$$

Thus, (4.9) and (4.12) yield

$$|\eta(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_1} - (\mathbf{U}^T\bar{\mathbf{y}})_{i_1}| < \epsilon, \tag{4.13}$$

$$|\eta(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_2} - (\mathbf{U}^T\bar{\mathbf{y}})_{i_2}| < \epsilon. \tag{4.14}$$

Note that in addition, (4.5) implies that

$$(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_1}(\mathbf{U}^T\bar{\mathbf{y}})_{i_1} \geq 0, \tag{4.15}$$

$$(\mathbf{U}^T\mathbf{x}^{k_{t_1}^1})_{i_2}(\mathbf{U}^T\bar{\mathbf{y}})_{i_2} \geq 0. \tag{4.16}$$ ∎

Combining (4.10), (4.13) and (4.15) implies that $\eta < 1$. This is a direct consequence of the following simple result on real numbers: if $a, b \in \mathbb{R}$ satisfy (for some $\epsilon > 0$) $ab \geq 0, |a| - |b| > \epsilon, |\eta a - b| < \epsilon$, then $\eta < 1$. Similarly, (4.11), (4.14) and (4.16) imply that $\eta > 1$. We thus obtained a contradiction to our assumption that $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$.

Remark 4.6. Convergence of the sequence generated by the projected gradient method can be established using the Kurdyka-Łojasiewicz property under the condition that the step-sizes t_k reside in the interval $(2\|\mathbf{A}\|_2, M)$ for some $M > 2\|\mathbf{A}\|_2$, see [2] and the precise statement in [7, Proposition 3].

4.3. Convergence to the Optimal Solution – the "Easy Case". So far, we have established the convergence of the sequence generated by an FOCM, but we did not prove that it converges to the global optimal solution of problem (TRS). This is actually not correct in general. What we can prove is that both in the easy and hard cases, certain choices of the initial vector $\mathbf{x}^0$ will guarantee convergence to the optimal solution.

Theorem 4.7 below shows that in the setting of the "easy case", if $\mathbf{x}^0 \in S_E$, then an FOCM will converge to the optimal solution of problem (TRS). This is done by showing that the limit of the sequence is a stationary point in S_E, which by the sufficient condition described in Theorem 3.1, implies that the limit is an optimal solution.

THEOREM 4.7 (**convergence to the optimal solution in the "easy case"**). *Let $\{\mathbf{x}^k\}_{k=0}^{\infty}$ be a sequence generated by an FOCM. If $\mathbf{x}^0 \in S_E$, then the sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ converges to the optimal solution of problem (TRS).*

Proof. By the definition of an FOCM, for any $k \in \mathbb{N}$ there exists $\boldsymbol{\theta}^k \in \mathbb{R}^2_+ \setminus \{\mathbf{0}\}$ such that

$$\mathbf{x}^{k+1} = \theta_1^k \mathbf{x}^k + \theta_2^k \left(-\nabla q(\mathbf{x}^k)\right),$$

and $\mathbf{x}^{k+1} \in B$. We will show that the whole sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ is in S_E by induction. By the statement of the theorem, $\mathbf{x}^0 \in S_E$. Assume that $\mathbf{x}^k \in S_E$. Then for any $i \in I_-$,

$$\begin{aligned}\text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)\left(-\mathbf{U}^T\nabla q(\mathbf{x}^k)\right)_i &= \text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)\left(-2d_i(\mathbf{U}^T\mathbf{x}^k)_i + 2(\mathbf{U}^T\mathbf{b})_i\right)\\ &= -2d_i\text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)(\mathbf{U}^T\mathbf{x}^k)_i + 2|(\mathbf{U}^T\mathbf{b})_i|\\ &\geq 0.\end{aligned}$$

where the inequality follows by the inclusion $\mathbf{x}^k \in S_E$ and the fact that $d_i \leq 0$ for any $i \in I_-$.

We have thus shown that if $\mathbf{x}^k \in S_E$, then $-\nabla q(\mathbf{x}^k) \in S_E$. This result and the fact that $\boldsymbol{\theta}^k \in \mathbb{R}^2_+ \setminus \{\mathbf{0}\}$ imply that for any $i \in I_-$,

$$\begin{aligned}&\text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)(\mathbf{U}^T\mathbf{x}^{k+1})_i =\\ &\qquad \theta_1^k\text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)(\mathbf{U}^T\mathbf{x}^k)_i + \theta_2^k\text{sgn}\left((\mathbf{U}^T\mathbf{b})_i\right)\left(-\mathbf{U}^T\nabla q(\mathbf{x}^k)\right)_i \geq 0. \quad \square\end{aligned}$$

Thus, $\mathbf{x}^{k+1} \in S_E$, showing that $\mathbf{x}^k \in S_E$ for any $k \in \mathbb{N}$. Therefore, since $\{\mathbf{x}^k\}_{k=0}^{\infty}$ converges to some stationary point $\bar{\mathbf{x}}$ (Theorem 4.5) and since S_E is a closed set, it follows that $\bar{\mathbf{x}} \in S_E$. We have thus shown that the limit point $\bar{\mathbf{x}}$ of the the sequence is a stationary point in S_E, and thus, by Theorem 3.1, it is an optimal solution of problem (TRS).

The set S_E as defined in (3.1) is given in terms of the spectral decomposition of $\mathbf{A}$. Thus, the task of finding a point $\mathbf{x}^0 \in S_E$ seems to require some spectral knowledge; however, such knowledge is not really needed by the trivial observation that $\mathbf{0} \in S_E$. Consequently, Theorem 4.7 ensures that in the "easy case", if an FOCM is initialized with $\mathbf{x}^0 = \mathbf{0}$, then it is guaranteed to converge to the optimal solution.

COROLLARY 4.8. *Suppose that the "easy case" holds and let $\{\mathbf{x}^k\}_{k=0}^{\infty}$ be a sequence generated by an FOCM initialized with $\mathbf{x}^0 = \mathbf{0}$. Then the sequence converges to an optimal solution of problem (TRS).*

4.4. Convergence to an Optimal Solution – the "Hard Case". In the "hard case", we will obtain a result that resembles a well-known result (see for example [13, Section 8.2.1]) for problem (TRS) with $\mathbf{b} = \mathbf{0}$, which shows that the so-called "power iteration method" converges to the dominant eigenvector of the matrix $\mathbf{A}$ under the condition that the dominant eigenvalue is unique[3] and the mild condition that the initial point $\mathbf{x}^0$ satisfies

$$\mathbf{x}^0 \not\perp \mathcal{E}_1. \tag{4.17}$$

The above condition is the same as the relation $\mathbf{x}^0 \in S_H$. We note that the power iteration method is known to be equivalent to the conditional gradient method (see

[3] This condition can be relaxed to some extent [30, Chapter 9].

[20]), and as such, it is an FOCM. Theorem 4.9 below shows that if the "hard case" holds and $\mathbf{b} \neq \mathbf{0}$, then under the same mild condition on the initial vector (4.17), convergence of any FOCM to the optimal solution of problem (TRS) is warranted. The proof is based on the sufficient optimality condition given in Theorem 3.3.

THEOREM 4.9 (**convergence to an optimal solution in the "hard case").** *Suppose that the "hard case" holds and that* $\mathbf{b} \neq \mathbf{0}$. *Let* $\{\mathbf{x}^k\}_{k=0}^{\infty}$ *be a sequence generated by an FOCM with* $\mathbf{x}^0 \in S_H$. *Then the sequence* $\{\mathbf{x}^k\}_{k=0}^{\infty}$ *converges to an optimal solution of problem* (TRS).

Proof. By Theorem 4.5 $\{\mathbf{x}^k\}_{k=0}^{\infty}$ converges, and we will denote its limit by $\bar{\mathbf{x}}$. By the definition of an FOCM, $\bar{\mathbf{x}}$ is a stationary point of problem (TRS). As usual, the associated Lagrange multiplier of $\bar{\mathbf{x}}$ will be denoted by $\lambda_{(\bar{\mathbf{x}})}$. By Theorem 2.5(a), $(\bar{\mathbf{x}}, \lambda_{(\bar{\mathbf{x}})})$ satisfies the equations

$$(\lambda_{(\bar{\mathbf{x}})} + d_i)(\mathbf{U}^T\bar{\mathbf{x}})_i = (\mathbf{U}^T\mathbf{b})_i, \quad i = 1, 2, \ldots, n. \tag{4.18}$$

We will prove that $\bar{\mathbf{x}}$ is an optimal solution of (TRS). If $d_1 \geq 0$, then problem (TRS) is convex, and hence $\bar{\mathbf{x}}$ must be an optimal solution, and we are done. We will henceforth consider the case where $d_1 < 0$.

Assume by contradiction that $\bar{\mathbf{x}}$ is not an optimal solution of problem (TRS). By Theorem 3.3, this entails that $\bar{\mathbf{x}} \notin S_H$, which by the definition of S_H (see (3.2)) implies that $(\mathbf{U}^T\bar{\mathbf{x}})_i = 0$ for any $i \in E_1$, and in particular that

$$(\mathbf{U}^T\bar{\mathbf{x}})_1 = 0. \tag{4.19}$$

Another consequence of the fact that $\bar{\mathbf{x}}$ is a non-optimal stationary point of problem (TRS) and Theorem 2.5(b) is that

$$\lambda_{(\bar{\mathbf{x}})} < -d_1. \tag{4.20}$$

Since $\mathbf{x}^0 \in S_H$, it follows that there exists $i \in E_1$ for which $(\mathbf{U}^T\mathbf{x}^0)_i \neq 0$ and we assume w.l.o.g. that $(\mathbf{U}^T\mathbf{x}^0)_1 \neq 0$.

Since $(\mathbf{U}^T\mathbf{b})_1 = 0$, the update formula of the method (4.1) can be written as

$$(\mathbf{U}^T\mathbf{x}^{k+1})_1 = \theta_1^k(\mathbf{U}^T\mathbf{x}^k)_1 + 2\theta_2^k(-d_1(\mathbf{U}^T\mathbf{x}^k)_1) = [\theta_1^k - 2\theta_2^k d_1](\mathbf{U}^T\mathbf{x}^k)_1,$$

meaning that

$$(\mathbf{U}^T\mathbf{x}^{k+1})_1 = \alpha_k(\mathbf{U}^T\mathbf{x}^k)_1, \tag{4.21}$$

where $\alpha_k = \theta_1^k - 2\theta_2^k d_1$. Note that $\alpha_k > 0$ since $d_1 < 0$ and $\boldsymbol{\theta}^k \in \mathbb{R}^2_+ \setminus \{\mathbf{0}\}$ for all k. Combining the fact that $(\mathbf{U}^T\mathbf{x}^0)_1 \neq 0$ and the validity of (4.21) with $\alpha_k > 0$, we can conclude that

$$(\mathbf{U}^T\mathbf{x}^k)_1 \neq 0 \text{ for any } k \in \mathbb{N}. \tag{4.22}$$

Since $\mathbf{b} \neq \mathbf{0}$, there exists $i \in \{1, 2, \ldots, n\}$ such that $(\mathbf{U}^T\mathbf{b})_i \neq 0$, which by (4.18) implies that $\lambda_{(\bar{\mathbf{x}})} \neq -d_i$ and

$$(\mathbf{U}^T\bar{\mathbf{x}})_i \neq 0. \tag{4.23}$$

Since $(\mathbf{U}^T\mathbf{x}^k)_i \to (\mathbf{U}^T\bar{\mathbf{x}})_i$ as $k \to \infty$, it follows by (4.18) and (4.23) that there exists $K_1 \in \mathbb{N}$ such that

$$(\mathbf{U}^T\mathbf{x}^k)_i[(\lambda_{(\bar{\mathbf{x}})} + d_i)(\mathbf{U}^T\mathbf{b})_i] > 0 \text{ for all } k \geq K_1. \tag{4.24}$$

The last equation implies further that

$$(\mathbf{U}^T\mathbf{x}^k)_i(\mathbf{U}^T\mathbf{x}^{k+1})_i > 0 \text{ for all } k \geq K_1. \tag{4.25}$$

For any $k \geq K_1$, we can write

$$(\mathbf{U}^T\mathbf{x}^{k+1})_i = \theta_1^k(\mathbf{U}^T\mathbf{x}^k)_i + 2\theta_2^k\left(-d_i(\mathbf{U}^T\mathbf{x}^k)_i + (\mathbf{U}^T\mathbf{b})_i\right),$$

meaning that

$$(\mathbf{U}^T\mathbf{x}^{k+1})_i = \eta_k(\mathbf{U}^T\mathbf{x}^k)_i, \tag{4.26}$$

where

$$\eta_k = \theta_1^k + 2\theta_2^k\left(-d_i + \frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\mathbf{x}^k)_i}\right).$$

Combining (4.25) and (4.26) we can also conclude that $\eta_k > 0$ for all $k \geq K_1$.

We now consider two cases:

(i) $\boldsymbol{\lambda_{(\bar{\mathbf{x}})} < -d_i}$. In this case, by (4.24), it follows that $\frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\mathbf{x}^k)_i} < 0$ for all $k \geq K_1$. We can therefore conclude that for any $k \geq K_1$

$$0 \leq \eta_k = \theta_1^k + 2\theta_2^k\left(-d_i + \frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\mathbf{x}^k)_i}\right) \leq \theta_1^k + 2\theta_2^k(-d_i) \leq \theta_1^k + 2\theta_2^k(-d_1) = \alpha_k.$$

(ii) $\boldsymbol{\lambda_{(\bar{\mathbf{x}})} > -d_i}$. In this case, by (4.18), it follows that

$$|(\mathbf{U}^T\bar{\mathbf{x}})_i| = \frac{|(\mathbf{U}^T\mathbf{b})_i|}{\lambda_{(\bar{\mathbf{x}})} + d_i} > \frac{|(\mathbf{U}^T\mathbf{b})_i|}{-d_1 + d_i},$$

where the inequality is due to (4.20). The above inequality can be also rewritten as

$$-d_1 > -d_i + \frac{|(\mathbf{U}^T\mathbf{b})_i|}{|(\mathbf{U}^T\bar{\mathbf{x}})_i|} = -d_i + \frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\bar{\mathbf{x}})_i},$$

where the equality is due to (4.18). Since $(\mathbf{U}^T\mathbf{x}^k)_i \to (\mathbf{U}^T\bar{\mathbf{x}})_i$ as $k \to \infty$ we obtain that there is some K_2 such that

$$-d_1 > -d_i + \frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\mathbf{x}^k)_i} \text{ for all } k \geq K_2.$$

We can thus conclude that in this case, for any $k \geq \max\{K_1, K_2\}$, it holds that

$$0 \leq \eta_k = \theta_1^k + 2\theta_2^k\left(-d_i + \frac{(\mathbf{U}^T\mathbf{b})_i}{(\mathbf{U}^T\mathbf{x}^k)_i}\right) \leq \theta_1^k + 2\theta_2^k(-d_1) = \alpha_k.$$

In both cases, we showed that there exists $K \in \mathbb{N}$ ($K = K_1$ in the first case and $K = \max\{K_1, K_2\}$ in the second case) such that for all $k \geq K$ it holds that $0 \leq \eta_k \leq \alpha_k$. We can now conclude that for any $k \geq K$,

$$\begin{aligned}
|(\mathbf{U}^T\mathbf{x}^{k+1})_i| &= \textstyle\prod_{j=K}^k \eta_j |(\mathbf{U}^T\mathbf{x}^K)_i| && [(4.26)] \\
&\leq \textstyle\prod_{j=K}^k \alpha_j |(\mathbf{U}^T\mathbf{x}^K)_i| && [0 \leq \eta_k \leq \alpha_k] \\
&= \textstyle\prod_{j=K}^k \alpha_j |(\mathbf{U}^T\mathbf{x}^K)_1| \frac{|(\mathbf{U}^T\mathbf{x}^K)_i|}{|(\mathbf{U}^T\mathbf{x}^K)_1|} && [(4.22)] \\
&= |(\mathbf{U}^T\mathbf{x}^k)_1| \frac{|(\mathbf{U}^T\mathbf{x}^K)_i|}{|(\mathbf{U}^T\mathbf{x}^K)_1|} && [(4.21)].
\end{aligned}$$

Since $\lim_{k\to\infty}(\mathbf{U}^T\mathbf{x}^k)_1 = (\mathbf{U}^T\bar{\mathbf{x}})_1 = 0$ (equation (4.19)), it follows by taking $k \to \infty$ in the above inequality that $(\mathbf{U}^T\bar{\mathbf{x}})_i = 0$, which is a contradiction to (4.23), thus establishing the desired result that $\bar{\mathbf{x}}$ is an optimal solution of problem (TRS). ∎

Remark 4.10 (**finding a point satisfying $\mathbf{x} \in \boldsymbol{S_H}$**). The condition $\mathbf{x}^0 \in S_H$, which is also the same as condition (4.17), can be more explicitly written as

$$(\mathbf{U}^T\mathbf{x}^0)_i \neq 0 \text{ for some } i \in E_1.$$

This is an extremely mild condition in the sense that if $\mathbf{x}^0$ is randomly generated via a continuous distribution with support B (for example, uniform distribution over B), then the probability that it will be satisfied is 1, and in this case, an FOCM will surely converge to the optimal solution of problem (TRS), as guaranteed by Theorem 4.9.

4.5. A Double-Start FOCM Method. Verifying whether the "easy" or "hard" cases hold requires spectral information which we do not assume to posses, and thus in this common case of uncertainty, we can employ an FOCM twice with two starting points: $\mathbf{x}^0 = \mathbf{0}$ and $\mathbf{x}^0$ which is randomly generated via a continuous distribution whose support is B. The best of the two resulting points (in terms of the objective function q) will be an optimal solution of problem (TRS) in probability 1.

Double-Start FOCM

Given an FOCM, execute the following:

(a) Employ the FOCM with initial point $\mathbf{x}^0 = \mathbf{0}$ and obtain an output $\bar{\mathbf{x}}$.
(b) Employ the FOCM with initial point $\mathbf{x}^0$ chosen via a continuous distribution function over B and obtain an output $\tilde{\mathbf{x}}$.
(c) The output of the method is $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{z}}\{q(\mathbf{z}) : \mathbf{z} \in \{\bar{\mathbf{x}}, \tilde{\mathbf{x}}\}\}$.

4.6. A Numerical Example. The GP and CG methods, belonging to the FOCM class, are attractive candidates for solving problem (TRS). Besides their extreme ease of implementation, they exhibit a low computational cost per iteration, which amounts to the matrix-vector product involved in the gradient calculation. Thus, both methods are applicable for solving large scale instances of problem (TRS), especially when the matrix $\mathbf{A}$ is sparse. In order to further motivate the benefits of solving problem (TRS) by an FOCM, we illustrate in the following example the potential efficiency of the GP and CG methods.

We considered 30 realizations of the model data $(\mathbf{A}, \mathbf{b})$ with dimension $n = 1000$ generated as follows. Each entry of the vectors $\mathbf{x}^*, \mathbf{u}, \in \mathbb{R}^n$ was first sampled uniformly in the interval $[-0.5, 0.5]$ and then both vectors were normalized to have unit length. The entries of the vector $\mathbf{d} \in \mathbb{R}^n$ were first sampled uniformly in the interval $[-5, 5]$; then, $\mathbf{d}$ was sorted in a non-decreasing order and its first element was set to -5. We then formed the matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where $\mathbf{D} = \operatorname{diag}(\mathbf{d})$ and $\mathbf{U} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$. Finally, we sampled $\lambda_{(\mathbf{x}^*)}$ uniformly in the interval $[5, 10]$ and set $\mathbf{b} = \mathbf{U}(\mathbf{D} + \lambda_{(\mathbf{x}^*)}\mathbf{I})\mathbf{U}^T\mathbf{x}^*$. This way, by Theorem 2.5, $\mathbf{x}^*$ is guaranteed to be the optimal solution. We made sure that each model satisfies the "easy case" assumption. Similar settings were previously considered in [23, 27].

Each method was initialized with the zeros vector and terminated after 1000 iterations. We used $\bar{L} = \|\mathbf{A}\|_\infty$ for the constant step-size selection rule of the GP method (obviously, $\bar{L} \geq \|\mathbf{A}\|_2$). For GP with a backtracking step-size strategy we used the parameters $\gamma = \eta^{-1} = 0.4$ and $s = \|\mathbf{A}\|_\infty$. This numerical illustration was conducted

on a PC with a 3.60GHz processor and 16GB RAM.
Table 1 shows the overall average running time in seconds and the average number of iterations needed in order to obtain a solution with objective value within tolerance $\tau \in \{10^{-6}, 10^{-10}, 10^{-14}\}$ of the optimal value $q_* = q(\mathbf{x}^*)$. Figure 1 exhibits the average relative optimality gap, which is defined as $(q(\mathbf{x}^k) - q_*)/|q_*|$, for the first 30 iterations.

	CG (exact)	GP (backtracking)	GP (constant)
time (sec.)	0.321	0.331	0.315
$\tau = 10^{-6}$	49	63	135
$\tau = 10^{-10}$	149	154	423
$\tau = 10^{-14}$	254	247	726

Table 1: *Average running time and number of iterations required in order to obtain a solution* $\mathbf{x}^k$ *such that* $q(\mathbf{x}^k) - q_* \leq \tau$.

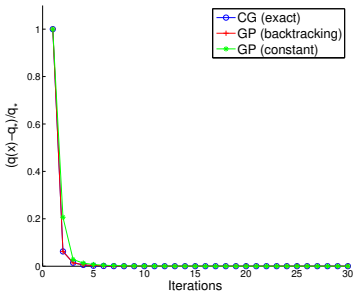

Fig. 1: *Average relative optimality gap for the first 30 iterations.*

It is evident that the sequences generated by the GP (constant), GP (backtracking) and CG (exact) exhibit a rapid convergence of the objective values where the first slightly lags behind the latter two.

REFERENCES

[1] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda, *Solving the trust-region subproblem by a generalized eigenvalue problem*, SIAM J. Optim., 27 (2017), pp. 269–291.
[2] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods*, Math. Program., 137 (2013), pp. 91–129.
[3] A. Beck, *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*, vol. 19, SIAM, 2014.
[4] A. Beck, *First-Order Methods in Optimization*, vol. 25, SIAM, 2017.
[5] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Program., 72 (1996), pp. 51–63.
[6] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2nd ed., Sept. 1999.
[7] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.
[8] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*, vol. 1, SIAM, 2000.
[9] J. B. Erway, P. E. Gill, and J. D. Griffin, *Iterative methods for finding a trust-region step*, SIAM J. Optim., 20 (2009), pp. 1110–1131.
[10] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.
[11] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114 (1989), pp. 815–839.
[12] D. M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Comput., 2 (1981), pp. 186–197.
[13] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.
[14] N. I. Gould, S. Lucidi, M. Roma, and P. L. Toint, *Solving the trust-region subproblem using the lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[15] W. W. Hager, *Minimizing a quadratic over a sphere*, SIAM J. Optim., 12 (2001), pp. 188–208.

[16] E. Hazan and T. Koren, *A linear-time algorithm for trust region problems*, Math. Program., 158 (2016), pp. 363–381.
[17] N. Ho-Nguyen and F. Kilinc-Karzan, *A second-order cone based approach for solving the trust-region subproblem and its variants*, SIAM J. Optim., 27 (2017), pp. 1485–1512.
[18] J. Lampe, M. Rojas, D. C. Sorensen, and H. Voss, *Accelerating the lstrs algorithm*, SIAM J. Sci. Comput., 33 (2011), pp. 175–194.
[19] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Comput. Math. Math. Phys., 6 (1966), pp. 787–823.
[20] R. Luss and M. Teboulle, *Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint*, SIAM Rev., 55 (2013), pp. 65–98.
[21] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Comput., 4 (1983), pp. 553–572.
[22] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Math. Program., 77 (1997), pp. 273–299.
[23] M. Rojas, S. A. Santos, and D. C. Sorensen, *A new matrix-free algorithm for the large-scale trust-region subproblem*, SIAM J. Optim., 11 (2001), pp. 611–646.
[24] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
[25] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.
[26] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[27] P. D. Tao and L. T. H. An, *A dc optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.
[28] P. L. Toint, *Towards an efficient sparsity exploiting newton method for minimization*, in Sparse Matrices and Their Uses, Academic Press, London, New York, 1981, pp. 57–88.
[29] J. Wang and Y. Xia, *A linear-time algorithm for the trust region subproblem based on hidden convexity*, Optim. Lett., (2016), pp. 1–8.
[30] J. H. Wilkinson, *The algebraic eigenvalue problem*, vol. 87, Clarendon Press Oxford, 1965.
[31] Y. Ye, *A new complexity result on minimization of a quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, Princeton University Press, 1992, pp. 19–31.