Inexact Variable Metric Method for Convex-Constrained Optimization Problems

Douglas S. Gonçalves * Max L. N. Gonçalves † Tiago C. Menezes†

September 17, 2020

Abstract

This paper is concerned with the inexact variable metric method for solving convex-constrained optimization problems. At each iteration of this method, the search direction is obtained by inexactly minimizing a strictly convex quadratic function over the closed convex feasible set. Here, we propose a new inexactness criterion for the search direction subproblems. Under mild assumptions, we prove that any accumulation point of the sequence generated by the new method is a stationary point of the problem under consideration. In order to illustrate the practical advantages of the new approach, we report some numerical experiments. In particular, we present an application where our concept of the inexact solutions is quite appealing.

Keywords: Convex-constrained optimization problem; approximate solution; projected gradient method; spectral gradient method; inexact variable metric method.

1 Introduction

This paper considers the following convex-constrained optimization problem

$$\min_{x\in C} f(x), \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function and $C \subseteq \mathbb{R}^n$ is a nonempty convex closed set. Recall that x_* is a stationary point of (1) iff $\langle \nabla f(x_*), y-x_*\rangle \geq 0$, for all $y\in C$. Owing to its efficiency and low memory requirements, the spectral projected gradient (SPG) method proposed in [5] has been considered

*Departamento de Matemática, Universidade Federal de Santa Catarina, Florianópolis, SC 88040-900, Brazil. (E-mail: douglas@mtm.ufsc.br. The work of this author was supported in part by CNPq Grant 421386/2016-9.

†IME, Universidade Federal de Goiás, Goiânia, GO 74001-970, Brazil. (E-mails: maxlng@ufg.br and tiagmenezes1@gmail.com). The work of these authors was supported in part by CAPES, FAPEG/PRONEM-201710267000532, and CNPq Grants 302666/2017-6 and 408123/2018-4.

an efficient approach for solving large-scale convex-constrained optimization problems. Given an arbitrary initial point $x_0 \in C$, the SPG method generates a sequence of iterates by the rule

$$x_{k+1} = x_k + \alpha_k d_k, \quad k \geq 0, \tag{2}$$

where the stepsize α_k is obtained by the nonmonotone line search strategies proposed in [19] and the search direction d_k is defined as

$$d_k = P_C(x_k - (1/\lambda_k)\nabla f(x_k)) - x_k,$$

where P_C denotes the orthogonal projection on C and λ_k is the Barzilai-Borwein scaling [4] defined by

$$\lambda_0 \in [\lambda_{min}, \lambda_{max}], \qquad \lambda_k = \min\{\lambda_{max}, \max\{\lambda_{min}, a_k/b_k\}\}, \tag{3}$$

with $0 < \lambda_{min} < \lambda_{max}$, $b_k := \langle x_k - x_{k-1}, x_k - x_{k-1}\rangle$ and $a_k := \langle x_k - x_{k-1}, \nabla f(x_k) - \nabla f(x_{k-1})\rangle$. The convergence results and/or numerical experiments illustrating the practical behavior of the SPG method were discussed in [5] and in many subsequent works including [3, 6, 8, 10, 21, 24, 25, 28, 15, 30, 31].

It is well-known that depending on the geometry of C, the orthogonal projection onto it neither has a closed-form nor can be easily computed. For this reason, paper [7] (see also [2]) proposed an inexact version of the SPG method in which approximate projections are allowed. Indeed, a more general approach, called Inexact Variable Metric method (IVM), was proposed. It differs from the SPG method by the fact that the search direction d_k in (2) is computed such that $x_k + d_k \in C$ and

$$Q_k(d_k) \leq \eta Q_k(\bar{d}_k), \tag{4}$$

where $\eta \in (0,1]$,

$$\bar{d}_k := \operatorname{argmin}_{x_k+d\in C} Q_k(d) := \frac{1}{2}\langle d, B_k d\rangle + \langle \nabla f(x_k), d\rangle, \tag{5}$$

and $B_k \in \mathbb{R}^{n\times n}$ is a suitable symmetric positive definite matrix. If $B_k := \lambda_k I$ for every $k \geq 0$, where λ_k is as in (3), the inexact variable metric method corresponds to an inexact version of the SPG method. Note also that if $\eta = 1$, the inexact variable metric method reduces to its exact version. It is not hard to verify that $\bar{d}_k$ in (5) is equivalent to $\bar{d}_k = \bar{y}_k - x_k$, with

$$\bar{y}_k = \operatorname{argmin}_{y\in C}\ q_k(y) := \frac{1}{2}\langle B_k y, y\rangle + \langle \nabla f(x_k) - B_k x_k, y\rangle. \tag{6}$$

In its turn, $\bar{y}_k$ in (6) is equivalent to

$$\bar{y}_k := \operatorname{argmin}_{y\in C} \frac{1}{2}\|y - (x_k - B_k^{-1}\nabla f(x_k))\|^2_{B_k}, \tag{7}$$

where $\|\cdot\|^2_{B_k} := \langle B_k \cdot, \cdot\rangle$. Therefore, d_k in (4) can also be interpreted as an approximation of the search direction $\bar{d}_k$ of the projected (in the norm $\|\cdot\|_{B_k}$) quasi-Newton method.

At first sight, a drawback of the inexact criterion in (4) is that it requires the optimal value of the problem in (5). It was presented in [2, 7] some applications in which it is possible to establish a sequence of lower bounds $C_l \leq Q_k(\bar{d}_k)$ that converges to the value $Q_k(\bar{d}_k)$ as l goes to infinity. Hence, criterion (4) is satisfied when the verifiable condition $Q_k(d_k) \leq \eta C_l$ holds. Specifically, by considering the inexact SPG method and assuming that C is a finite intersection of closed and convex easy sets (i.e., where the orthogonal projection onto each of them can be easily computed), paper [7] used a modification of the Dykstra projection method (see [9]), which generates the desired sequence $\{C_l\}$, to inexactly solve the problem in (5). In [2], it was assumed that $C = \{x \in \mathbb{R}^n : Ax \leq b\}$, where $A \in \mathbb{R}^{m\times n}$, and a dual approach was used to compute the sequence $\{C_l\}$. The associated dual subproblems in the latter reference were solved by means of an active-set box-constraint quadratic optimizer with a proximal-point type unconstrained algorithm for minimization within the current faces. It is not clear however how the strategies in [2, 7] can be employed or even how an inexact direction satisfying (4) can be obtained for other complex feasible sets (where the projection cannot be easily performed).

Therefore, the goal of this paper is to study an inexact variable metric method with a different inexactness criterion for the subproblems (6). We first present a concept of approximate solution for the problem in (6), which does not require the knowledge of its optimal value. The new criterion can be verified by finding the infimum of a linear function over the feasible set C. Such verification comes for free when the conditional gradient method (Frank-Wolfe) [11, 12] is used to solve the problem in (6). Under mild assumptions, we prove that any accumulation point of the sequence generated by the proposed method is a stationary point of (1). In order to illustrate the practical advantages of the new approach, we report some numerical experiments. In particular, we present an application where our concept of inexact solutions is quite appealing.

The organization of the paper is as follows. Section 2 presents our concept of approximate solution of the problem in (6) and the associated inexact variable metric method. The global convergence of the method is also discussed in Section 2. Some numerical experiments of the proposed method are presented in Section 3 and final remarks are given in Section 4.

2 The inexact variable metric method and its convergence analysis

In this section, we propose and study an inexact variable metric method for solving (1). Basically, the method differs from the one studied in [2, 7] by using a different inaccuracy criterion for the search direction subproblems.

Let $\mathbb{B}$ be the set of $n \times n$ symmetric positive definite matrices such that

$$\|B\| \leq L \qquad \text{and} \qquad \|B^{-1}\| \leq L, \tag{8}$$

where $L > 1$ and $\|.\|$ is a sub-multiplicative matrix norm. Note that $\mathbb{B}$ is a compact set of $\mathbb{R}^{n\times n}$. Consider also the inner product on $\mathbb{R}^n$ defined by $\langle x, z\rangle_B = \langle x, Bz\rangle$, where $B \in \mathbb{B}$ and $\langle ., .\rangle$ denotes the usual inner product. Notice that the corresponding induced norm $\|.\|_B$ is equivalent to the Euclidean norm on $\mathbb{R}^n$, since the following inequalities hold

$$\frac{1}{\|B^{-1}\|}\|x\|^2 \leq \|x\|_B^2 \leq \|B\|\|x\|^2. \tag{9}$$

We next define our concept of approximate solution for (6).

Definition 1. *Given $B \in \mathbb{B}$, $x \in C$ and $\varepsilon \geq 0$, we say that $\tilde{y}_B(x)$ is an ε–approximate solution for the problem*

$$\min_{y\in C}\ q(y) := \frac{1}{2}\langle By, y\rangle + \langle \nabla f(x) - Bx, y\rangle \tag{10}$$

iff

$$\tilde{y}_B(x) \in C \qquad \textit{and} \qquad \langle B(x - \tilde{y}_B(x)) - \nabla f(x), y - \tilde{y}_B(x)\rangle \leq \varepsilon, \quad \forall\, y \in C. \tag{11}$$

Remark 1. *Since in* (10) *we are minimizing a strictly convex quadratic function over a convex set, condition* (11) *is a natural condition for an approximate solution. Indeed, the optimality condition for* (10) *is $\langle \nabla q(\bar{y}), y - \bar{y}\rangle \geq 0, \forall y \in C$. Hence, one could define an approximate solution as $\tilde{y} \in C$ such that $\langle \nabla q(\tilde{y}), y - \tilde{y}\rangle \geq -\varepsilon, \forall y \in C$, which coincides with* (11).

Note that, if $\tilde{y}_B(x)$ is a *zero*–approximate solution, then $\langle \nabla q(\tilde{y}_B(x)), y - \tilde{y}_B(x)\rangle \geq 0$, for all $y \in C$, which in turn implies that $\tilde{y}_B(x)$ is the unique exact solution of (10). We will denote this exact solution by $y_B(x)$. Since $y_B(x) \in C$ and $\tilde{y}_B(x) \in C$, for $B \in \mathbb{B}$, $x \in \mathbb{R}^n$ and $\varepsilon \geq 0$, it follows from Definition 1 that

$$\langle B\tilde{y}_B(x) - Bx + \nabla f(x), \tilde{y}_B(x) - y_B(x)\rangle \leq \varepsilon, \qquad \langle Bx - \nabla f(x) - By_B(x), \tilde{y}_B(x) - y_B(x)\rangle \leq 0.$$

By adding the last two inequalities, we obtain the following relationship between y_B and $\tilde{y}_B$:

$$\|\tilde{y}_B(x) - y_B(x)\|_B \leq \sqrt{\varepsilon}. \tag{12}$$

Moreover, since problem (10) can be rewritten, unless constant terms, as

$$\min_{y\in C} \frac{1}{2}\|y - (x - B^{-1}\nabla f(x))\|_B^2, \tag{13}$$

and (11) is equivalent to

$$\langle x - B^{-1}\nabla f(x) - \tilde{y}_B(x), y - \tilde{y}_B(x)\rangle_B \leq \varepsilon, \quad \forall\, y \in C,$$

where $\tilde{y}_B(x) \in C$, we can say that $\tilde{y}_B(x)$ is an approximate projection (in the norm $\|\cdot\|_B$) of an unconstrained quasi-Newton step.

We emphasize that our criterion can be easily checked when, for example, C is bounded and the conditional gradient method [11] is used to solve (10). The conditional gradient (CondG) method, also known as Frank-Wolfe method [12], is designed to solve the convex optimization problem $\min_{x\in\Omega} h(x)$, where Ω is a nonempty compact convex set and h is a differentiable convex function. Given $z_{j-1} \in \Omega$, its j-th step first finds $\bar{z}_j$ as a minimum of the linear function $\langle \nabla h(z_{j-1}), \cdot \rangle$ over Ω and then set $z_j = (1-\alpha_j)z_{j-1} + \alpha_j \bar{z}_j$ for some $\alpha_j \in [0,1]$. Its major distinguishing feature compared to other first-order algorithms such as the projected gradient (or accelerated gradient) method is that it replaces the usual projection onto Ω by a linear oracle which computes $\bar{z}_j$ as above. Since, for some relevant cases of Ω (for example, when Ω is the spectrahedron; see Section 3.2), the latter operation is considerably cheaper than the first one, the CondG method is competitive with first-order projection methods and it has recently re-gained attention in different application areas (see, e.g., [13, 23]). If we apply the CondG method to (10), then $\bar{z}_j$ is a solution of the subproblem

$$\begin{aligned} \min \quad & \langle B(z_{j-1}-x) + \nabla f(x), z - z_{j-1} \rangle, \\ \text{s.t.} \quad & z \in C \end{aligned} \tag{14}$$

and, hence, if the CondG iterations are stopped when

$$\langle B(z_{j-1}-x) + \nabla f(x), \bar{z}_j - z_{j-1} \rangle \geq -\varepsilon, \tag{15}$$

then condition (11) holds with $\tilde{y}_B(x) = z_{j-1}$.

We next discuss a way to use the CondG method to obtain an approximate solution for (10) when the diameter of C is very large or even when C is unbounded. Note that the exact solution $y_B(x)$ of (10) satisfies

$$\langle B(x - y_B(x)) - \nabla f(x), x - y_B(x) \rangle \leq 0,$$

which, combined with Cauchy-Schwarz inequality yields

$$\|x - y_B(x)\|_B^2 \leq \langle \nabla f(x), x - y_B(x) \rangle \leq \|\nabla f(x)\| \|x - y_B(x)\|.$$

It follows from the last inequality and (9) that

$$\|x - y_B(x)\| \leq \|B^{-1}\| \|\nabla f(x)\|,$$

which implies that the ball $\mathcal{B}\left(x, \|B^{-1}\|\|\nabla f(x)\|\right)$ contains the (unknown) exact solution $y_B(x)$ of (10). Therefore, one can apply the conditional gradient method to (10) with C replaced by $C \cap \mathcal{B}\left(x, \|B^{-1}\|\|\nabla f(x)\|\right)$ in order to obtain a point $\tilde{y}_B(x)$ satisfying

$$\tilde{y}_B(x) \in C, \quad \langle B(x - \tilde{y}_B(x)) - \nabla f(x), y - \tilde{y}_B(x) \rangle \leq \varepsilon, \quad \forall\, y \in C \cap \mathcal{B}\left(x, \|B^{-1}\|\|\nabla f(x)\|\right). \tag{16}$$

It can be proven, using that the function $q(y)$ in (10) is strongly convex and $y_B(x) \in C \cap \mathcal{B}\left(x, \|B^{-1}\|\|\nabla f(x)\|\right)$, that if $\varepsilon = 0$ in the last inequality, then $\tilde{y}_B(x) = y_B(x)$. Therefore, we claim that the well-definiteness as well as the convergence results of the proposed algorithm can also be shown if (11) is replaced by (16).

We also mention that other iterative methods can take place to obtain an ε–approximate solution for (10), being enough to solve periodically, or at each iteration j, the linear subproblem (14) to test our criterion: unboundness of the linear subproblem implies that the criterion does not hold.

We are now able to formally describe the inexact method for solving (1).

Algorithm 1 (New IVM) Given $x_0 \in C$, $B_0 \in \mathbb{B}$, $\tau \in (0,1)$, an integer $M \geq 1$ and $\{\theta_k\} \subset [0,\infty)$.

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

1. **(Inexact search direction)** Compute $d_k = \tilde{y}_{B_k}(x_k) - x_k$, where $\tilde{y}_{B_k}(x_k) \in C$ and

$$\langle B_k(x_k - \tilde{y}_{B_k}(x_k)) - \nabla f(x_k), y - \tilde{y}_{B_k}(x_k)\rangle \leq \varepsilon_k := \theta_k^2 \|\tilde{y}_{B_k}(x_k) - x_k\|_{B_k}^2, \quad \forall\, y \in C, \tag{17}$$

   i.e., $\tilde{y}_{B_k}(x_k)$ is an ε_k–approximate solution of (6).

2. **(Termination Criterion)** If $\|d_k\| = 0$, then **stop**.

3. **(Backtracking)** Define $f_{max} = \max\{f(x_{k-j}); 0 \leq j \leq \min\{k, M-1\}\}$. Set $\alpha \leftarrow 1$.

   3.1. If

$$f(x_k + \alpha d_k) \leq f_{max} + \tau\alpha\langle \nabla f(x_k), d_k\rangle, \tag{18}$$

   then $\alpha_k = \alpha$, $x_{k+1} = x_k + \alpha d_k$, and go to Step 4. Otherwise, set $\alpha \leftarrow \alpha/2$ and go to Step 3.1.

4. **(Update of the Hessian approximation)** Form a matrix $B_{k+1} \in \mathbb{B}$.

Remark 2. *(i) If $d_k = 0$, then $\tilde{y}_{B_k}(x_k) = x_k$. Hence, it follows from* (17) *that*

$$\langle \nabla f(x_k), y - x_k\rangle \geq 0, \quad \forall y \in C,$$

i.e., $x_k \in C$ is a stationary point of (1)*. Conversely, if x_k is a stationary point of* (1)*, then it follows from* (17) *with $y = x_k$ and the optimality condition that $\tilde{y}_{B_k}(x_k) = x_k$. (ii) Notice that Step 1 is well-defined because the exact solution of* (6) *clearly satisfies* (17)*. Nevertheless, iterative methods can be used to obtain an approximate solution of* (6) *such that condition* (17) *holds. If the conditional gradient is employed, for example, the stopping criterion* (15) *now reads*

$$\langle B_k(z_{j-1} - x_k) + \nabla f(x_k), \bar{z}_j - z_{j-1}\rangle \geq -\theta_k^2 \|z_{j-1} - x_k\|_{B_k}^2. \tag{19}$$

From item (i) in this remark, we observe that if $z_{j-1} = x_k$, then

$$\forall z \in C : \langle \nabla f(x_k), z - x_k\rangle = \langle \nabla f(x_k), z - z_{j-1}\rangle \geq \langle \nabla f(x_k), \bar{z}_j - z_{j-1}\rangle \geq 0,$$

showing that z_{j-1} is stationary for the original problem. (iii) If $\theta_k = 0$ in (17)*, we obtain that $\tilde{y}_{B_k}(x_k)$ is the unique exact solution of the problem* (6) *and then, the inexact variable metric method reduces to its exact version. Additionally, if $B_k := \lambda_k I$ for every $k \geq 0$, where λ_k is as in* (3)*, the inexact variable metric method corresponds to the inexact SPG method. (iv) As it will be proven later, the search directions generated by Algorithm 1 are descent directions, which will imply that the backtracking process given in step 3 is well-*

defined. (v) There are different choices for, or ways to build, the matrix B_k. For example, B_k can be the Hessian of function f if it is positive definite or a modification of it in order to guarantee the positiveness of the approximation. The approximate B_k can be a specific multiple of the identity matrix such as the spectral choice in [5, 7].

In order to investigate the global convergence of the method, we need to establish some properties of its search directions.

Proposition 1. *Assume that the sequence $\{\theta_k\}$ satisfies $\theta_k \leq \bar{\theta}$ for all $k \geq 0$, where $\bar{\theta} \in [0,1)$. Then, for every $k \geq 0$, we have*

$$\langle d_k, \nabla f(x_k)\rangle \leq -(1-\bar{\theta}^2)L\|d_k\|^2 \tag{20}$$

and

$$\frac{1}{(1+\bar{\theta})L}\|y_{B_k}(x_k)-x_k\| \leq \|d_k\| \leq \frac{L}{1-\bar{\theta}^2}\|\nabla f(x_k)\|, \tag{21}$$

where $y_{B_k}(x_k)$ is the exact solution of the problem (6).

Proof. Since $d_k = \tilde{y}_{B_k}(x_k) - x_k$, from (17) with $y = x_k$, we have

$$\langle \nabla f(x_k), d_k\rangle \leq (\theta_k^2 - 1)\|d_k\|_{B_k}^2, \tag{22}$$

which, combined with the fact that $\theta_k \leq \bar{\theta} < 1$ for all $k \geq 0$, (8) and (9), yields

$$\langle \nabla f(x_k), d_k\rangle \leq -(1-\bar{\theta}^2)L\|d_k\|^2.$$

Thus, (20) is proved. It follows from (22) and the Cauchy-Schwarz inequality that

$$(1-\theta_k^2)\|d_k\|_{B_k}^2 \leq -\langle \nabla f(x_k), d_k\rangle \leq \|\nabla f(x_k)\|\|d_k\|.$$

Hence, the second inequality in (21) now follows from (8), (9) and the fact that $\theta_k \leq \bar{\theta} < 1$ for all $k \geq 0$. Now, from (9) and the triangle inequality, we obtain

$$\begin{aligned}
\|y_{B_k}(x_k) - x_k\| &\leq \|B_k^{-1}\|^{1/2}\|y_{B_k}(x_k) - x_k\|_{B_k} \\
&\leq \|B_k^{-1}\|^{1/2}\|y_{B_k}(x_k) - \tilde{y}_{B_k}(x_k)\|_{B_k} + \|B_k^{-1}\|^{1/2}\|\tilde{y}_{B_k}(x_k) - x_k\|_{B_k} \\
&\leq \|B_k^{-1}\|^{1/2}[\sqrt{\varepsilon_k} + \|d_k\|_{B_k}],
\end{aligned}$$

where last inequality is due to (12) and $d_k = \tilde{y}_{B_k}(x_k) - x_k$. Since $\varepsilon_k = \theta_k^2\|d_k\|_{B_k}^2$ (see Step 1 of the Algorithm 1), it follows from the last inequality that

$$\|y_{B_k}(x_k) - x_k\| \leq (1+\theta_k)\|B_k^{-1}\|^{1/2}\|d_k\|_{B_k}.$$

Therefore, the first inequality in (21) now follows from (8), (9) and the fact that $\theta_k \leq \bar{\theta}$ for all $k \geq 0$.

□

We next establish the global convergence of the Algorithm 1.

Theorem 2. *Assume that the level set $C_0 := \{x \in C : f(x) \leq f(x_0)\}$ is bounded and the sequence $\{\theta_k\}$ satisfies $\theta_k \leq \bar{\theta}$ for all $k \geq 0$, where $\bar{\theta} \in [0,1)$. Then, either the Algorithm 1 stops at some stationary point x_k, or every limit point of the generated sequence is stationary.*

Proof. If Algorithm 1 stops at a point x_k, then $d_k = 0$. Hence, $\tilde{y}_{B_k}(x_k) = x_k$ and it follows from (17) that

$$\langle \nabla f(x_k), y - x_k \rangle \geq 0, \quad \forall y \in C,$$

i.e., x_k is a stationary point of (1). If $d_k \neq 0$, for every $k \geq 0$, it follows from (20) that d_k is a descent direction. So, the backtracking process given in step 3 is well-defined, and, as a consequence, the Algorithm 1 is also well-defined. Our goal is now to show that every limit point of the $\{x_k\}$ is a stationary point of (1). Let $l(k)$ be an integer such that $k - \min\{k, M-1\} \leq l(k) \leq k$ and

$$f(x_{l(k)}) = \max_{0 \leq j \leq \min\{k, M-1\}} f(x_{k-j}).$$

Using the first part of the proof of the theorem in [19] with $m(k) := \min\{k, M-1\}$ (note that this choice of $m(k)$ satisfies the conditions of the mentioned theorem), it can be shown that $\{f(x_{l(k)})\}$ is monotonically nonincreasing, and from the boundness of C_0 we have that $\{f(x_{l(k)})\}$ admits a limit for $k \to \infty$. From (18), it follows, for $k > M-1$, that

$$f(x_{l(k)}) \leq f(x_{l(l(k)-1)}) + \tau\alpha_{(l(k)-1)} \langle \nabla f(x_{(l(k)-1)}), d_{(l(k)-1)} \rangle. \tag{23}$$

Now, since $\alpha_{(l(k)-1)} > 0$ and $\langle \nabla f(x_{(l(k)-1)})), d_{(l(k)-1)} \rangle < 0$, by taking limits in (23), it follows that

$$\lim_{k \to \infty} \alpha_{(l(k)-1)} \langle \nabla f(x_{(l(k)-1)}), d_{(l(k)-1)} \rangle = 0.$$

Moreover, from (20) and (21), we conclude that

$$\lim_{k \to \infty} \alpha_{(l(k)-1)} \| y_{B_{(l(k)-1)}}(x_{(l(k)-1)}) - x_{(l(k)-1)} \|^2 = 0,$$

and following the idea in the proof of in the theorem of [19], we can write

$$\lim_{k \to \infty} \alpha_k \| y_{B_k}(x_k) - x_k \|^2 = 0. \tag{24}$$

Let $x_* \in C$ be a limit point of $\{x_k\}$. Relabel $\{x_k\}$ a subsequence converging to x_*. From (24), there exists a subsequence of indices $K_1 \subset K$ such that: (i) $\lim_{k \in K_1} \| y_{B_k}(x_k) - x_k \| = 0$ or (ii) $\lim_{k \in K_1} \alpha_k = 0$.
(i) By the compactness of $\mathbb{B}$ we can extract a subsequence of indices such $K_2 \subset K_1$ such that

$$\lim_{k \in K_2} B_k = B_* \in \mathbb{B}.$$

Hence, we obtain, by continuity of $y_B(x)$ (see Lemma 3 in the appendix), that $\|y_{B_*}(x_*) - x_*\| = 0$, or equivalently, $y_{B_*}(x_*) = x_*$, which, from the definition $y_{B_*}(x_*)$ (see Definition 1), implies that

$$\langle \nabla f(x_*), y - x_* \rangle \geq 0, \quad \forall y \in C,$$

i.e., x_* is a stationary point of (1).
(ii) Let α_k be the step chosen in the Step 3.2 such that $\alpha_k = \bar{\alpha}_k/2$, where $\bar{\alpha}_k$ was the last step that failed in (18), i.e.

$$f(x_k + \bar{\alpha}_k d_k) > \max_{0 \leq j \leq \min\{k, M-1\}} f(x_{k-j}) + \tau \bar{\alpha}_k \langle \nabla f(x_k), d_k \rangle \geq f(x_k) + \tau \bar{\alpha}_k \langle \nabla f(x_k), d_k \rangle. \tag{25}$$

Now define $s_k = \bar{\alpha}_k d_k$. By the mean value theorem, there exists $\mu_k \in [0,1]$ such that the relation in (25) can be written as

$$\langle \nabla f(x_k + \mu_k s_k), s_k \rangle = f(x_k + s_k) - f(x_k) > \tau \langle \nabla f(x_k), s_k \rangle. \tag{26}$$

On the other hand, as $\{x_k\}$ is bounded and f has continuous derivatives, we have, by (21), that $\{d_k\}$ is bounded. Thus, since $s_k = 2\alpha_k d_k$, and $\lim_{k \in K_1} \alpha_k = 0$, we obtain that s_k goes to zero as $k \in K_1$ goes to infinity. So, from (26), we have

$$\langle \nabla f(x_k + \mu_k s_k), \frac{s_k}{\|s_k\|} \rangle > \tau \langle \nabla f(x_k), \frac{s_k}{\|s_k\|} \rangle. \tag{27}$$

By taking limit in the last inequality as $k \in K_3$ goes to infinity, where $K_3 \subset K_1$ is such that $lim_{k \in K_3} \{s_k/\|s_k\|\}$ converges to s, we obtain $(1-\tau)\langle \nabla f(x_*), s \rangle \geq 0$. Since $(1-\tau) > 0$, we have

$$\langle \nabla f(x_*), s \rangle \geq 0. \tag{28}$$

Now, as d_k is a descent direction for f at x_k (see (20)) and $s_k = \bar{\alpha}_k d_k$, we find

$$\langle \nabla f(x_k), \frac{s_k}{\|s_k\|} \rangle < 0.$$

Hence, $\langle \nabla f(x_*), s \rangle \leq 0$, which, combined with (28), implies that $\langle \nabla f(x_*), s \rangle = 0$. Using (20), (21) and the definition of s_k, we have

$$\langle \nabla f(x_k), \frac{s_k}{\|s_k\|} \rangle \leq -(1 - \bar{\theta}^2) L \|d_k\| \leq -(1 - \bar{\theta}) \|y_{B_k}(x_k) - x_k\|.$$

By the compactness of $\mathbb{B}$ we can extract a subsequence of indices such $K_4 \subset K_3$ such that $\lim_{k \in K_4} B_k = B_* \in \mathbb{B}$. Therefore, by taking limit in the last inequality as $k \in K_4$ goes to infinity, we have

$$0 = (1 + \bar{\theta}) L \langle \nabla f(x_*), s \rangle \leq -(1 - \bar{\theta}) \|y_{B_*}(x_*) - x_*\|.$$

Since $\bar{\theta} < 1$, we obtain $y_{B_*}(x_*) = x_*$, which, from the definition $y_{B_*}(x_*)$ (see Definition 1), implies that x_* is a stationary point of (1). $\square$

3 Numerical experiments

We split the numerical experiments in two sets. First, in Section 3.1, where polyhedral feasible sets $C = \{x \in \mathbb{R}^n \ : \ Ax \leq b\}$ are considered, we aim to evaluate the impact of the new inexactness criterion (17) of Algorithm 1 in comparison with the criterion (4) used in the Inexact SPG (ISPG) of [2]. Then, in Section 3.2, we consider a feasible set for which the use of inexact variable metric methods is quite appealing (because the cost of an exact solution of (6) is prohibitive) and we show that Algorithm 1 achieves good results with respect to its exact counterpart and an off-the-shelf solver.

All experiments were carried out in Matlab R2018b, in a laptop running Mac OS X 10.13.6, with 8GB of RAM and 1.8 Ghz Intel Core i5 processor.

We implemented Algorithm 1 with the following parameters: $\tau = 10^{-4}$, $\lambda_{\min} = 10^{-10}$, $\lambda_{\max} = 10^{10}$ and $\theta_k = \bar{\theta} = 0.9995$.

3.1 Polyhedral feasible set

In order to put in perspective the new inexactness criterion (17) with the previously proposed criterion (4), we consider a subset of the linearly constrained problems from the CUTEr collection [18] used in [2, Table 2] and compare the results with those obtained by Algorithm 1.

As in [2], the tolerance in the stopping criterion $\|d_k\| < \varepsilon$ was set to $\varepsilon = 10^{-6}$. For this set of experiments, we considered the variant of Algorithm 1 with $B_k = \lambda_k I$, with λ_k as in (3) (and $\lambda_0 = 1$).

Since the feasible set $C = \{x \in \mathbb{R}^n \ : \ Ax \leq b\}$ is described by linear inequality constraints, the IVM subproblems (10) are in fact (strictly convex) quadratic programming problems. Bound constraints were treated as ordinary inequality constraints. The subproblems (10) were solved by using a variant of the Frank-Wolfe algorithm including away-steps [20] whose subproblems (see Eq. (14)) were solved by the Matlab linear programming solver `linprog`. In order to handle problems with unbounded C, we have included an additional constraint corresponding to the ball of (16) in infinity norm, so that the subproblems (14) are well-defined.

Table 1 presents the number of variables n, number of original inequality constraints m, the number of outer (OUTIT) and inner (INNIT) iterations required by each method and the optimal value $f(x^*)$ found by Algorithm 1. The values in Table 1 for ISPG were collected from Table 2 in [2]. From these figures, we observe that Algorithm 1 requires less outer iterations, also reducing the number of inner iterations in 9 out of 14 problems. For some problems, fewer outer iterations came at the cost of more inner iterations. We remark that the problems SIPOW (see [26] for details) are in fact linear programming problems that would be solved in a single (outer) iteration by Algorithm 1 if $\lambda_0 = 0$. Nevertheless, we also observe a better performance of Algorithm 1 for problems where the solution is an extreme point of the feasible polyhedron, as in HS24, HS36 and HS44NEW.

			ISPG [2]		Algorithm 1		
Problem	n	m	OUTIT	INNIT	OUTIT	INNIT	$f(x^*)$
HS24	2	3	11	17	7	8	-1.000
HS35	3	1	17	15	12	22	0.1111
HS35I	3	1	17	15	12	22	0.1111
HS36	3	1	12	30	1	2	-3300.
HS37	3	2	28	35	14	67	-3456
HS44NEW	4	6	13	29	4	9	-15
HS76	4	3	14	19	8	66	-4.6818
HS76I	4	3	14	19	8	66	-4.6818
SIPOW1	2	2000	15	641	2	3	-1
SIPOW1M	2	2000	13	656	2	3	-1.0000
SIPOW2	2	2000	11	349	2	3	-1
SIPOW2M	2	2000	12	323	2	3	-1.0000
SIPOW3	4	2000	11	1355	2	3	0.5346
SIPOW4	4	2000	11	1398	2	4	0.2724

Table 1: Comparison of ISPG [2] and Algorithm 1 on CUTEr problems

3.2 Least squares on the spectrahedron

In this section, we consider the least squares problem over the spectrahedron:

$$\begin{aligned} \min_{X \in \mathbb{S}^n} \quad & \frac{1}{2}\|AX - Z\|_F^2 \\ \text{s.t.} \quad & \operatorname{tr}(X) = 1 \\ & X \succeq 0, \end{aligned} \tag{29}$$

where $A, Z \in \mathbb{R}^{m\times n}$, with $m > n$, $\mathbb{S}^n$ denotes the vector space of symmetric matrices of order n equipped with the trace inner product $\langle X,Y\rangle = \operatorname{tr}(XY)$ and induced norm $\|X\|_F^2 = \langle X,X\rangle$, and $X \succeq 0$ means that X is positive semidefinite.

Problem (29) is related to important applications in many areas. For example, in nonlinear optimization, it can be used to estimate positive definite approximations for the inverse Hessian in quasi-Newton methods, whereas in structural analysis it can be used to estimate the compliance matrix of an elastic structure (see [29] for details).

Clearly, the feasible set $C = \{X \in \mathbb{S}^n \,:\, \operatorname{tr}(X) = 1,\, X \succeq 0\}$ is convex and compact whereas the objective function of (29) is strictly convex, provided $\operatorname{rank}(A) = n$.

We remark that, for this feasible set, the computation of the exact orthogonal projection[1] of a point $Y \in \mathbb{S}^n$ onto C requires the full eigendecomposition of Y which is prohibitive for large values of n (for

[1] with respect to the Frobenius norm

details, see [22, 16]). Since the projection problem is equivalent to (13) when B is a positive multiple of the identity, and (13) in its turn is equivalent to (10) for any positive definite B, we expect that the cost of solving (10) *exactly* becomes also prohibitive for large dimensions. Therefore, it seems reasonable to consider inexact variable metric methods in this case.

Since C is neither polyhedral nor a finite intersection of easy convex sets, the approaches in [2, 7] are not directly applicable.

On the other hand, if an ε-approximate solution of (10) is allowed (in the sense of (11)), one could employ, for example, the Frank-Wolfe algorithm [12] whose iteration cost is dictated by an extreme eigenpair computation when C is the spectrahedron (see [17] and references therein). If only a few Frank-Wolfe iterations are required to achieve (11), then overall savings, in terms of computational effort, may be considerable when running variants of Algorithm 1.

To numerically investigate this claim, we consider random instances of problem (29) and compare the performance of variants of Algorithm 1 with SPG using exact projections [16] and an interior point method [27].

The first group of problems consists of dense small problems with $n < m \leq 1000$. The matrices A were randomly generated with entries sampled from a uniform distribution in the interval $[0,1]$. Then, given a positive integer q, we build a symmetric matrix $\tilde{X}$ with q eigenvalues equal to $1/q$, one equal to -1, and all others equal to zero. Finally, we set $Z = A\tilde{X}$. In general, this procedure results in nonzero residue problems.

For this group of problems, we consider two variants of Algorithm 1, namely, "Inexact Newton" where $B_k = A^T A$ and "Inexact SPG" where $B_k = \lambda_k I$, with λ_k as in (3), and compare them with the off-the-shelf solver QSDP [27] which implements an interior point method for convex quadratic semidefinite programming problems.

Since the classic Frank-Wolfe is known for its slow $O(1/\varepsilon)$ convergence [14], we consider a variant of the conditional gradient proposed in [1], that we shall call FW-p. FW-p is specialized for the spectrahedron and, by exploiting an estimate of the solution rank p, achieves $O(\kappa \log(1/\varepsilon))$ convergence rate, where κ is the condition number[2] of the subproblem (10). This scheme fits well the Inexact SPG because $B_k = \lambda_k I$ implies in $\kappa = 1$. Preliminary experiments revealed that it also works fine with Inexact Newton as long as B_k remains well-conditioned. Therefore, we adopt FW-p in our inexact methods except for Inexact Newton in the first group of problems (which contains some ill-conditioned ones) for which we use the classic Frank-Wolfe and limit the number of inner iterations to 500.

The tolerance in the stopping criterion $\|d_k\| < \varepsilon$ was set to $\varepsilon = 10^{-3}$, as well as the tolerance for the duality gap in QSDP.

Concerning the parameter M of the nonmonotone line-search, we observed in preliminary numerical experiments that the full-step ($\alpha_k = 1$) was always accepted in the "Inexact Newton", so we kept $M = 1$ for this variant. For the "Inexact SPG", we did not observe a pronounced improvement for $M = 5$ or $M = 10$ for this test set, thus we decided to go on with the monotone line-search ($M = 1$).

For each problem, we consider 3 starting points given by $X_0(\gamma) = (1-\gamma)(1/n)I + \gamma\hat{X}$, where $\hat{X} = e_1 e_1^T$

[2]Assuming that the convex function q is α-strongly convex and β-smooth, the corresponding condition number is given by $\kappa = \beta/\alpha$. See [1] for details.

				QSDP			Inexact Newton			Inexact SPG		
n	m	q	γ	it.	time	$f(X_k)$	it.	time	$f(X_k)$	it	time	$f(X_k)$
10	100	4	0	9	1.24	9.13	7	1.57	9.14	12	**0.23**	9.13
			0.5	9	0.67	9.13	7	0.96	9.14	17	**0.21**	9.13
			0.99	9	0.55	9.13	7	0.91	9.14	16	**0.18**	9.13
50	200	4	0	16	2.54	16.51	8	4.75	16.52	38	**0.70**	16.51
			0.5	16	1.98	16.51	8	3.68	16.52	34	**0.57**	16.51
			0.99	16	1.54	16.51	9	3.74	16.52	44	**0.72**	16.51
50	200	10	0	14	1.54	16.70	10	6.29	16.71	28	**0.56**	16.70
			0.5	14	1.43	16.70	11	5.64	16.71	34	**0.57**	16.71
			0.99	14	1.30	16.70	11	5.15	16.71	48	**0.78**	16.70
100	400	5	0	18	4.11	29.07	8	6.27	29.07	42	**0.93**	29.04
			0.5	18	3.23	29.07	9	6.17	29.07	54	**1.03**	29.06
			0.99	18	3.07	29.07	10	6.12	29.07	55	**1.17**	29.07
200	800	5	0	23	11.04	53.02	10	13.84	53.04	74	**2.72**	53.01
			0.5	23	11.31	53.02	10	12.68	53.04	48	**2.21**	53.20
			0.99	23	11.38	53.02	11	12.81	53.04	54	**2.16**	53.11
200	800	20	0	20	10.53	52.60	15	26.33	52.62	34	**1.08**	52.64
			0.5	20	8.84	52.60	15	24.86	52.62	37	**1.75**	53.03
			0.99	20	8.83	52.60	16	25.12	52.62	46	**1.83**	53.06
400	1000	5	0	28	74.01	65.31	10	41.39	65.35	43	**4.57**	66.28
			0.5	28	72.10	65.31	11	43.76	65.34	60	**8.65**	66.14
			0.99	28	73.05	65.31	11	40.70	65.34	71	**9.13**	66.19

Table 2: Numerical results for dense small problems.

(e_1 is the first canonical vector) and $\gamma \in \{0, 0.5, 0.99\}$.

Table 2 brings the number of iterations, running time in seconds, and the achieved objective value $f(X_k)$. The smallest running time for each problem is highlighted in bold. From these results, we observe that the Inexact SPG variant of Algorithm 1 provides a non-negligible speed-up with respect to QSDP in all problems. We point out that the number of inner iterations of ISPG was less than 50 in all problems. For Inexact Newton (IN), the times were greater than QSDP in all problems but the last. We also notice that the maximum number of 500 inner iterations was reached in the final iterations of IN for all problems.

We remark that the approximate optimal values do not match in some problems: a consequence of different stopping criteria. Nevertheless, due to the speed-up of ISPG, we can afford to be more restrictive in the stopping criterion. For example, if we decrease the tolerance of Inexact SPG to $\varepsilon = 10^{-4}$ in the problem defined by $n = 400$, $m = 1000$, $q = 5$ and $\gamma = 0$, it obtains $f(X_k) = 65.28$ after 96 iterations in 12 seconds.

In the second group of problems, we consider sparse matrices with dimensions $m > n \geq 1000$. The matrix A was build using the command `sprand(m,n,1e-4)` from Matlab, and $\tilde{X} = QDQ^T$, where Q is the product of a few Givens rotation matrices and D is a diagonal matrix with q entries equal to one and all others equal to zero. This ensures that $Z = A\tilde{X}$ is also sparse.

For this second test set, the interior point solver QSDP was left out of comparison due to excessively high running times. We replace it by a version of SPG where the projection is computed "exactly" as in [16]. This version is referred in Table 3 as "Exact SPG".

From Table 3, we observe that the Inexact SPG surpassed SPG with exact projections in the majority of problems. Inexact Newton also shows a good performance and becomes faster than Exact SPG as n and m increases.

4 Final remark

We proposed an inexact variable metric method (IVM), with a new inexactness criterion for its subproblems, for solving convex-constrained optimization problems. When necessary, such inexact solutions of the subproblems can be obtained by using suitable iterative algorithms; for example, the conditional gradient method (Frank-Wolfe) [11, 12] and its variants. Under mild assumptions, we proved that any accumulation point of the sequence generated by the proposed method is a stationary point of (1). Preliminary numerical experiments showed that the new algorithm works well and compares favorably with a previous IVM on linearly constrained problems, and with its exact version and the interior point method in [27] for semidefinite least squares problems.

Acknowledgement

We are indebted to two anonymous referees whose comments helped to improve this paper.

References

[1] Allen-Zhu, Z., Hazan, E., Hu, W., and Li, Y. Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls. In *Advances in Neural Information Processing Systems* (2017), pp. 6191–6200.

[2] Andreani, R., Birgin, E. G., Martínez, J. M., and Yuan, J. Spectral projected gradient and variable metric methods for optimization with linear inequalities. *IMA J. Numer. Anal. 25* (2005), 221–252.

[3] Andretta, M., Birgin, E. G., and Martínez, J. M. Partial spectral projected gradient method with active-set strategy for linearly constrained optimization. *Numer. Algorithms 53* (2009), 23–52.

[4] Barzilai, J., and Borwein, J. M. Two-point step size gradient methods. *IMA J. Numer. Anal. 8*, 1 (1988), 141–148.

				Exact SPG			Inexact Newton			Inexact SPG		
n	m	q	γ	it.	time	$f(X_k)$	it.	time	$f(X_k)$	it	time	$f(X_k)$
1000	2000	10	0.0	5	**2.03**	0.0794	4	6.44	0.0794	6	2.24	0.0794
			0.5	10	3.28	0.0794	4	5.77	0.0794	8	**2.08**	0.0794
			1.0	7	2.39	0.0794	4	4.69	0.0794	10	**2.15**	0.0794
1000	2000	20	0.0	6	**2.34**	0.2311	4	7.43	0.2311	6	2.79	0.2311
			0.5	7	2.48	0.2311	4	7.00	0.2311	8	**2.41**	0.2311
			1.0	8	2.73	0.2311	3	3.86	0.2311	8	**2.42**	0.2311
2000	4000	10	0.0	5	**12.33**	0.2304	3	15.73	0.2304	6	23.01	0.2304
			0.5	6	14.83	0.2304	3	8.28	0.2304	6	**6.99**	0.2304
			1.0	5	12.20	0.2304	3	**8.25**	0.2304	9	8.53	0.2304
2000	4000	20	0.0	5	12.22	0.9442	3	9.46	0.9442	5	**7.90**	0.9442
			0.5	5	12.04	0.9442	3	9.17	0.9442	6	**8.52**	0.9442
			1.0	6	20.92	0.9442	3	9.07	0.9442	7	**7.93**	0.9442
3000	6000	10	0.0	5	37.87	0.1377	3	21.00	0.1377	6	**17.60**	0.1377
			0.5	5	37.82	0.1377	3	20.56	0.1377	6	**15.95**	0.1377
			1.0	7	49.08	0.1377	3	21.26	0.1377	6	**15.27**	0.1377
3000	6000	20	0.0	3	24.95	1.0832	3	18.59	1.0833	5	**17.58**	1.0832
			0.5	4	30.90	1.0832	3	18.44	1.0833	5	**15.83**	1.0832
			1.0	4	30.38	1.0832	3	**18.27**	1.0833	6	18.32	1.0832
4000	8000	10	0.0	5	85.50	0.6192	3	34.17	0.6193	5	**28.83**	0.6192
			0.5	4	109	0.6192	3	33.60	0.6193	5	**28.20**	0.6192
			1.0	5	112	0.6192	3	35.93	0.6193	6	**31.29**	0.6192
4000	8000	20	0.0	5	117	2.9650	3	59.02	2.9650	5	**34.41**	0.2965
			0.5	5	116	2.9650	3	56.32	2.9650	5	**34.93**	2.9650
			1.0	6	144	2.9650	3	55.51	2.9650	6	**34.81**	2.9651
5000	10000	10	0.0	6	258	0.9923	3	64.83	0.9923	6	**57.97**	0.9923
			0.5	7	296	0.9923	3	64.69	0.9923	6	**59.79**	0.9923
			1.0	7	300	0.9923	3	**64.68**	0.9923	7	65.79	0.9923
5000	10000	20	0.0	3	107	2.9388	3	101	2.9388	4	**84.15**	2.9388
			0.5	4	138	2.9388	3	100	2.9388	4	**67.55**	2.9390
			1.0	4	139	2.9388	3	**78.33**	2.9390	6	82.42	2.9388

Table 3: Numerical results for sparse medium-scale problems.

[5] BIRGIN, E. G., MARTÍNEZ, J. M., AND RAYDAN, M. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim. 10* (2000), 1196–1211.

[6] BIRGIN, E. G., MARTÍNEZ, J. M., AND RAYDAN, M. Algorithm 813: SPG - software for convex-constrained optimization. *ACM Trans. Math. Software 27*, 3 (2001), 340–349.

[7] BIRGIN, E. G., MARTÍNEZ, J. M., AND RAYDAN, M. Inexact spectral projected gradient methods on convex sets. *IMA J. Numer Anal. 23* (2003), 539–559.

[8] BIRGIN, E. G., MARTÍNEZ, J. M., AND RAYDAN, M. Spectral projected gradient methods: Review and perspectives. *J. Stat. Softw. 60*, 3 (2014), 1–21.

[9] BOYLE, J. P., AND DYKSTRA, R. L. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference* (New York, NY, 1986), R. Dykstra, T. Robertson, and F. T. Wright, Eds., Springer New York, pp. 28–47.

[10] CORES, D., ESCALANTE, R., GONZÁLEZ-LIMA, M., AND JIMENEZ, O. On the use of the Spectral Projected Gradient method for Support Vector Machines. *J. Comput. Appl. Math. 28* (2009), 327 – 364.

[11] DUNN, J. C. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM J. Control. Optim. 18*, 5 (1980), 473–487.

[12] FRANK, M., AND WOLFE, P. An algorithm for quadratic programming. *Nav. Res. Logist. Q. 3*, 1-2 (1956), 95–110.

[13] FREUND, R., AND GRIGAS, P. New analysis and results for the Frank-Wolfe method. *Math. Program.* (2014), 1–32.

[14] GARBER, D. Faster projection-free convex optimization over the spectrahedron. In *Advances in Neural Information Processing Systems* (2016), pp. 874–882.

[15] GOMES-RUGGIERO, M. A., MARTÍNEZ, J. M., AND SANTOS, S. A. Spectral projected gradient method with inexact restoration for minimization with nonconvex constraints. *J. Sci. Comput. 31*, 3 (2009), 1628–1652.

[16] GONÇALVES, D., GOMES-RUGGIERO, M., AND LAVOR, C. A projected gradient method for optimization over density matrices. *Optim. Methods Softw. 31*, 2 (2016), 328–341.

[17] GONÇALVES, D. S., GONÇALVES, M. L. N., AND OLIVEIRA, F. R. Levenberg-Marquardt methods with inexact projections for constrained nonlinear systems. *arXiv e-prints* (Aug 2019), arXiv:1908.06118.

[18] GOULD, N. I. M., ORBAN, D., AND TOINT, P. L. Cuter, a constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Softw. 29* (2003), 373–394.

[19] GRIPPO, L., LAMPARIELLO, F., AND LUCIDI, S. A nonmonotone line search technique for Newton's method. *SIAM J. Numer. Anal. 23*, 4 (1986), 707–716.

[20] GUÉLAT, J., AND MARCOTTE, P. Some comments on Wolfe's 'away step'. *Math. Program. 35* (1986), 110–119.

[21] GUERRERO, J., RAYDAN, M., AND ROJAS, M. A hybrid-optimization method for large-scale non-negative full regularization in image restoration. *Inver. Probl. Sci. En. 21* (2011), 741–766.

[22] HIGHAM, N. J. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl. 103* (1988), 103 – 118.

[23] JAGGI, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013), vol. 28, pp. 427–435.

[24] LIU, J., AND DUAN, Y. Two spectral gradient projection methods for constrained equations and their linear convergence rate. *J. Inequal. Appl. 2015*, 1 (2015), 8.

[25] MARTÍNEZ, J. M., PILOTTA, E. A., AND RAYDAN, M. Spectral gradient methods for linearly constrained optimization. *J. Optim. Theory Appl. 125* (2005), 629–651.

[26] POWELL, M. J. D. Log barrier methods for semi-infinite programming calculations. In *Advances on Computer Mathematics and its Applications* (1993), E. A. Lipitakis, Ed., pp. 1–21.

[27] TOH, K.-C. An inexact primal–dual path following algorithm for convex quadratic SDP. *Math. Program. 112*, 1 (2008), 221–254.

[28] WANG, C., LIU, Q., AND YANG, X. Convergence properties of nonmonotone spectral projected gradient methods. *J. Comput. Appl. Math. 182* (2005), 51–66.

[29] WOODGATE, K. G. Least-squares solution of $F = PG$ over positive semidefinite symmetric P. *Linear Algebra Appl. 245* (1996), 171–190.

[30] YU, Z., LIN, J., SUN, J., XIAO, Y., LIU, L., AND LI, Z. Spectral gradient projection method for monotone nonlinear equations with convex constraints. *Appl. Numer. Math. 59*, 10 (2009), 2416 – 2423.

[31] ZHANG, L., AND ZHOU, W. Spectral gradient projection method for solving nonlinear monotone equations. *J. Comput. Appl. Math. 196*, 2 (2006), 478 – 484.

A Auxiliary lemma

Lemma 3. *The exact solution $y_B(x)$ of* (10) *is a continuous function of $B \in \mathbb{B}$ and $x \in C$.*

Proof. Let $x,z \in C$ and $B,D \in \mathbb{B}$. It follows from Definition (1) that

$$\begin{aligned} \langle B(y_B(x)-x)+\nabla f(x), y_B(z)-y_B(x)\rangle &\geq 0, \\ \langle B(y_B(z)-z)+\nabla f(z), y_B(x)-y_B(z)\rangle &\geq 0. \end{aligned}$$

Summing the above inequalities

$$\langle B(y_B(x)-y_B(z))-B(x-z)+\nabla f(x)-\nabla f(z), y_B(z)-y_B(x)\rangle \geq 0,$$

and after some manipulation

$$-\|y_B(z)-y_B(x)\|_B^2+\langle B(z-x), y_B(z)-y_B(x)\rangle+\langle \nabla f(x)-\nabla f(z), y_B(z)-y_B(x)\rangle \geq 0.$$

Then,

$$\frac{1}{\|B^{-1}\|}\|y_B(z)-y_B(x)\|^2 \leq (\|B\|\|z-x\|+\|\nabla f(x)-\nabla f(z)\|)\,\|y_B(z)-y_B(x)\|,$$

where in the left hand side we used (9) and on the right hand side we used Cauchy-Schwarz inequality and consistency of matrix norm.

Supposing $y_B(z) \neq y_B(x)$, from the above inequality and Eq. (8), we arrive at

$$\|y_B(z)-y_B(x)\| \leq L\,(L\|z-x\|+\|\nabla f(x)-\nabla f(z)\|)\,. \tag{30}$$

Since inequality (30) is also valid when $y_B(z)=y_B(x)$, x,z and B were taken arbitrarily, the inequality is valid for all $x,z \in C$ and $B \in \mathbb{B}$.

Now, again from Definition (1), we have in particular

$$\begin{aligned} \langle B(y_B(z)-z)+\nabla f(z), y_D(z)-y_B(z)\rangle &\geq 0, \\ \langle D(y_D(z)-z)+\nabla f(z), y_B(z)-y_D(z)\rangle &\geq 0. \end{aligned}$$

Summing the above inequalities yields

$$\langle By_B(z)-Bz+Dz-Dy_D(z), y_D(z)-y_B(z)\rangle \geq 0,$$

or, equivalently (after some manipulation),

$$\langle B(y_B(z)-y_D(z))+(D-B)(z-y_D(z)), y_D(z)-y_B(z)\rangle \geq 0.$$

The above inequality leads to

$$\|y_D(z)-y_B(z)\|_B^2 \leq \langle (D-B)(z-y_D(z)), y_D(z)-y_B(z)\rangle.$$

Assuming $y_D(z) \neq y_B(z)$, invoking (9) and the Cauchy-Schwarz inequality, we obtain

$$\|y_D(z) - y_B(z)\| \leq \|B^{-1}\| \|z - y_D(z)\| \|D - B\| \leq L\|z - y_D(z)\| \|D - B\|, \tag{31}$$

where the last inequality follows from (8). On the other hand, from Definition (1), we also have

$$\langle D(y_D(z) - z) + \nabla f(z), z - y_D(z) \rangle \geq 0,$$

or, equivalently,

$$\frac{1}{\|D^{-1}\|} \|z - y_D(z)\| \leq \|z - y_D(z)\|_D \leq \|\nabla f(z)\| \leq \|\nabla f(x)\| + \|\nabla f(x) - \nabla f(z)\|.$$

Combining the last inequality with (31) and Eq. (8), we obtain

$$\|y_D(z) - y_B(z)\| \leq \|B^{-1}\| \|z - y_D(z)\| \|D - B\| \leq L^2 \left(\|\nabla f(x)\| + \|\nabla f(x) - \nabla f(z)\|\right) \|D - B\|, \tag{32}$$

Since (32) also holds when $y_D(z) = y_B(z)$, and because z, B, D were chosen arbitrarily, we conclude that it is valid for all $z \in C$ and $B, D \in \mathbb{B}$.

Finally, using (30), (32) and the triangle inequality, we find

$$\begin{aligned} \|y_B(x) - y_D(z)\| &\leq \|y_B(x) - y_B(z)\| + \|y_B(z) - y_D(z)\| \\ &\leq L^2\|z - x\| + L(1 + L\|D - B\|)\|\nabla f(x) - \nabla f(z)\| + L^2\|\nabla f(x)\| \|D - B\|, \end{aligned}$$

which, combined with the fact that ∇f is continuous and $\mathbb{B}$ is compact, implies that $y_B(x)$ is continuous as a function of $x \in C$ and $B \in \mathbb{B}$. □