Frank–Wolfe algorithm for DC optimization problem

R. Díaz Millán *   O. P. Ferreira †   J. Ugon ‡

September 1, 2023

Abstract

In the present paper, we formulate two versions of Frank–Wolfe algorithm or conditional gradient method to solve the DC optimization problem with an adaptive step size. The DC objective function consists of two components; the first is thought to be differentiable with a continuous Lipschitz gradient, while the second is only thought to be convex. The second version is based on the first and employs finite differences to approximate the gradient of the first component of the objective function. In contrast to past formulations that used the curvature/Lipschitz-type constant of the objective function, the step size computed does not require any constant associated with the components. For the first version, we established that the algorithm is well-defined of the algorithm and that every limit point of the generated sequence is a stationary point of the problem. We also introduce the class of weak-star-convex functions and show that, despite the fact that these functions are non-convex in general, the rate of convergence of the first version of the algorithm to minimize these functions is $\mathcal{O}(1/k)$. The finite difference used to approximate the gradient in the second version of the Frank-Wolfe algorithm is computed with the step-size adaptively updated using two previous iterations. Unlike previous applications of finite difference in the Frank-Wolfe algorithm, which provided approximate gradients with absolute error, the one used here provides us with a relative error, simplifying the algorithm analysis. In this case, we show that all limit points of the generated sequence for the second version of the Frank-Wolfe algorithm are stationary points for the problem under consideration, and we establish that the rate of convergence for the duality gap is $\mathcal{O}(1/\sqrt{k})$.

Keywords: Frank-Wolfe method; DC optimization problem; finite difference; weak-star-convex function.

AMS subject classification: 90C25, 90C60, 90C30, 65K05.

1 Introduction

The DC optimization method involves minimizing a function that can be represented as the difference between two convex functions. We are interested in finding a solution

*School of Information Technology, Deakin University, Melbourne, Australia, e-mail:`r.diazmillan@deakin.edu.au`.

†IME, Universidade Federal de Goiás, Goiânia, Brazil, e-mail:`orizon@ufg.br`.

‡School of Information Technology, Deakin University, Melbourne, Australia, e-mail:`j.ugon@deakin.edu.au`.

to a constrained DC optimization problem, where the constraint set $\mathcal{C} \subset \mathbb{R}^n$ is a convex and compact set, and $f := g - h$ where $g : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function, and $h : \mathbb{R}^n \to \mathbb{R}$ is a convex function possibly non-differentiable. To the best of our knowledge, DC optimization can be traced back to pioneering works such as [39, 40], which introduced the first algorithms for this problem. Since then, the DC optimization has attracted the attention of the mathematical programming community (see for example [6, 10, 12, 30, 41]), not only for its own sake but also because it is an abstract model for several families of practical optimization problems. Although we are not concerned with practical issues at this time, we emphasize that practical applications emerge whenever the natural structure of the problem is modelled as a DC optimization problem, such as sparse generalized eigenvalue problems [38], sparse optimization problems [18], facility location and clustering problems [33].

The Frank-Wolfe algorithm has a long history, dating back to Frank and Wolfe's work in the 1950s to minimize convex quadratic functions over compact polyhedral sets, see [14]. This method was generalized about ten years later to minimize convex differentiable functions with Lipschitz continuous gradients and compact constraint convex sets, see [31]. Since then, this method has attracted the attention of several researchers who work with continuous optimization, thus becoming also known as the conditional gradient method. One of the factors that explain the interest in this method is its simplicity and ease of implementation. In fact, each method iteration only requires access to a linear minimization oracle over a compact convex set. It is also worth noting that, due to the method's simplicity, it allows for low storage costs and ready exploration of separability and sparsity, making its application in large-scale problems quite attractive. It is interesting to note that the popularity of this method has increased significantly in recent years as a result of the emergence of several applications in machine learning, see [23, 27, 28]. For all of these reasons, several variants of this method have arisen throughout the years and new properties of it have been discovered, resulting in a large literature on it, papers dealing with this method include [3, 7, 8, 16, 17, 20, 25, 29, 32].

In the present paper, we formulate two versions of *Frank–Wolfe algorithm or conditional gradient method to solve DC optimization problem with an adaptive step size.* The second version is based on the first and employs finite differences to approximate the gradient ∇g. It is worth mentioning that the Frank-Wolfe algorithm has previously been formulated in the context of DC programming, see [24]. See also [43] which establishes theoretical connections between DC algorithms and convex-concave procedures with the Frank-Wolfe algorithm. In contrast to the previous study, which assume that the curvature/Lipschitz-type constant of $f = g - h$ is bounded from above, the analysis of the Frank-Wolfe method done here just assumes that the gradient of the first component g of f is Lipschitz continuous. While designing the methods, even if we assume that the gradient of g is Lipschitz continuous in both formulations, we will not compute the step size using the Lipschitz constant of ∇g, as in earlier formulations that employed the curvature/Lipschitz-type constant of the objective function f. The step size will be computed adaptively, based on an idea introduced in [2] (see also [4, 36]) that approximates the Lipschitz constant. It has been shown in previous works that Frank-Wolfe algorithm produces a stationary point with a convergence rate of only $\mathcal{O}(1/\sqrt{k})$ for non-convex objective functions, see [26] (see also [24]). We introduce the concept of weak-star-convex function notion, which generalizes the star-convex function concept proposed in [34] as well as a few other related concepts, such as [21, 42]. The weak-star-convex objective functions are also taken into account in the convergence analysis of the first method presented here. The rate of convergence is proven to be $\mathcal{O}(1/k)$ for both the function values and the duality gap, despite the fact that the weak-star-convex functions are in general non-convex. As a result, among the functions for which the Frank–Wolfe method yields an approximate solution with a convergence rate of $\mathcal{O}(1/k)$, we include the set of weak-star-convex functions that are differences of convex functions with the first component having a Lipschitz gradient. As mentioned previously, the second version of the Frank-Wolfe algorithm uses finite difference to approximate the gradient of the function g. Similar to the strategies adopted in [19], the finite difference utilized here to approximate the gradient is computed with the step-size updated adaptively utilizing two previous iterations. Finite difference is an old concept that has been used to approximate derivatives in optimization settings (for example, see [35, Section 8.1]) and has already appeared in the study of the Frank-Wolfe algorithm (for example, see [13, 15, 23, 37]). It is worth noting that all previous applications of finite difference in the Frank-Wolfe algorithm provided approximate gradients with absolute error, whereas the one used here has the advantage of providing us with a relative error and thus being quite simple to analyze. We show that all limit points of the generated sequence for the second version of the Frank-Wolfe algorithm are stationary points for the problem under consideration, and we establish that the rate of convergence is $\mathcal{O}(1/\sqrt{k})$ for the duality gap.

The following describes the manner in which this paper is organized. Some notations and auxiliary results are presented in Section 2. Section 3 presents the problem, the hypotheses required, and some related notations. Section 4 is devoted to the formulation of the first version of the Frank-Wolfe algorithm, as well as its well definition and analysis of the sequence generated by it. Section 5 contains the formulation of the second version of the algorithm and its well definition and analysis. Finally, in Section 6, some conclusions are presented.

2 Preliminaries

In this section, we recall some notations, definitions and basics results used throughout the paper. A function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$, for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$. And φ is *strictly convex* when the last inequality is strict for $x \neq y$, for a comprehensive study of convex function see [22]. We say that $f : \mathbb{R}^n \to \mathbb{R}$ is *locally Lipschitz* if, for all $x \in \mathbb{R}^n$, there exist a constant $K_x > 0$ and a neighborhood U_x of x such that $|f(x) - f(y)| \leq K_x\|x - y\|$, for all $y \in U_x$. It is well known that, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then f is locally Lipschitz. And we also know that if $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable then f is locally Lipschitz, see [9, p. 32]. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. The *Clarke's subdifferential* of f at $x \in \mathbb{R}^n$ is given by $\partial_c f(x) = \{v \in \mathbb{R}^n : \ f^\circ(x; d) \geq v^{\mathrm{T}} d, \ \forall d \in \mathbb{R}^n\}$, where $f^\circ(x; d)$ is the *generalized directional derivative* of f at x in the direction d given by

$$f^\circ(x; d) = \limsup_{\substack{u \to x \\ t \downarrow 0}} \frac{f(u + td) - f(u)}{t}.$$

For an extensive study of locally Lipschitz functions and Clarke's subdifferential see [9, p. 27]. If f is convex, then $\partial_c f(x)$ coincides with the subdifferential $\partial f(x)$ in the sense of convex analysis, and $f^\circ(x;d)$ coincides with the usual directional derivative $f'(x;d)$; see [9, p. 36]. We recall that if $f:\mathbb{R}^n\to\mathbb{R}$ is differentiable, then $\partial_c f(x)=\{\nabla f(x)\}$ for any $x\in\mathbb{R}^n$; see [9, p. 33].

Theorem 2.1 ([9, p. 27])**.** *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a locally Lipschitz function. Then, $\partial_c f(x)$ is a nonempty, convex, compact subset of $\mathbb{R}^n$ and $\|v\|\le K_x$, for all $v\in\partial_c f(x)$, where $K_x>0$ is the Lipschitz constant of f around x. Moreover, $f^\circ(x;d)=\max\{v^{\mathrm{T}}d:\ v\in\partial_c f(x)\}$.*

Theorem 2.2 ([9, p. 38-39])**.** *Let $f:\mathbb{R}^n\to\mathbb{R}$ be given by $f=g-h$, where $g,h:\mathbb{R}^n\to\mathbb{R}$ is locally Lipschitz functions and g is differentiable. Then, $f^\circ(x;d)=\nabla g(x)^{\mathrm{T}}d-h'(x;d)$, for all $x,d\in\mathbb{R}^n$ and $\partial_c f(x)=\{\nabla g(x)\}-\partial h(x)$.*

The next result is a combination of Theorem 2.2 with [9, Corollary on p. 52].

Theorem 2.3. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a locally Lipschitz function and $\mathcal{C}\subset\mathbb{R}^n$ is a closed and convex set. If $x^*\in\mathcal{C}$ is a minimizer of f in $\mathcal{C}$, then $v^{\mathrm{T}}(x-x^*)\ge 0$, for all $v\in\partial_c f(x^*)$ and all $x\in\mathcal{C}$. As a consequence, if $f=g-h$ with $g,h:\mathbb{R}^n\to\mathbb{R}$ is locally Lipschitz functions and g is differentiable, then $(\nabla g(x^*)-u)^{\mathrm{T}}(x-x^*)\ge 0$, for all $u\in\partial_c h(x^*)$ and all $x\in\mathcal{C}$.*

Hence, according to Theorem 2.3, every point satisfying the following inequality $v^{\mathrm{T}}(x-x^*)\ge 0$, for all $v\in\partial_c f(x^*)$ and for all $x\in\mathcal{C}$, is called to be a *stationary point* of $\min_{x\in\mathcal{C}} f(x)$. A continuously differentiable function $g:\mathbb{R}^n\to\mathbb{R}$ has gradient ∇g is *L-Lipschitz continuous* on $\mathcal{C}\subset\mathbb{R}^n$ if there exists a Lipschitz constant $L>0$ such that $\|\nabla g(x)-\nabla g(y)\|\le L\|x-y\|$ for all $x,y\in\mathcal{C}$. Thus, by using the fundamental theorem of calculus, we obtain the following result whose proof can be found in [5, Proposition A.24], see also[11, Lemma 2.4.2].

Proposition 2.4. *Let $g:\mathbb{R}^n\to\mathbb{R}$ be a differentiable with gradient L-Lipschitz continuous on $\mathcal{C}\subset\mathbb{R}^n$, $x\in\mathcal{C}$, $v\in\mathbb{R}^n$ and $\lambda\in[0,1]$. If $x+\lambda v\in\mathcal{C}$, then $g(x+\lambda v)\le g(x)+\nabla g(x)^{\mathrm{T}}v\lambda+\frac{L}{2}\|v\|^2\lambda^2$.*

Proposition 2.5. *The function $h:\mathbb{R}^n\to\mathbb{R}$ is convex if and only if $h(y)\ge h(x)+\langle u,y-x\rangle$, for all $x,y\in\mathbb{R}^n$ and all $u\in\partial h(x)$.*

Proposition 2.6 ([22, Proposition 6.2.1])**.** *Let $h:\mathbb{R}^n\to\mathbb{R}$ be convex. Let $(x^k)_{k\in\mathbb{N}}$ and $(u^k)_{k\in\mathbb{N}}$ be sequences such that $u^k\in\partial h(x^k)$, for all $k\in\mathbb{N}$. If $\lim_{k\to+\infty}x^k=\bar{x}$ and $\lim_{k\to+\infty}u^k=\bar{u}$, then $\bar{u}\in\partial h(\bar{x})$.*

Proposition 2.7 ([22, Proposition 6.2.2])**.** *Let $h:\mathbb{R}^n\to\mathbb{R}$ be convex. The mapping ∂h is locally bounded, i.e. the image $\partial h(B)$ of a bounded set $B\subset\mathbb{R}^n$ is a bounded set in $\mathbb{R}^n$.*

In the following we recall an useful results for our study on iteration complexity bounds for Frank–Wolfe algorithm, its proof can be found in [1, Lemma 13.13, Ch. 13, p. 387].

Lemma 2.8. *Let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be nonnegative sequences of real numbers satisfying*

$$a_{k+1}\le a_k-b_k\beta_k+\frac{A}{2}\beta_k^2,\qquad k=0,1,2,\dots,$$

where $\beta_k=2/(k+2)$ and A is a positive number. Suppose that $a_k\le b_k$, for all k. Then *(i)* $a_k \leq \dfrac{2A}{k}$, *for all* $k = 1, 2, \ldots$.

(ii) $\displaystyle\min_{\ell \in \{\lfloor \frac{k}{2} \rfloor + 2, \ldots, k\}} b_\ell \leq \frac{8A}{k-2}$, *for all* $k = 3, 4, \ldots$, *where,* $\lfloor k/2 \rfloor = \max\{n \in \mathbb{N} : \ n \leq k/2\}$.

3 The DC optimization problem

We are interested in solving the following constrained DC optimization problem

$$\min_{x \in \mathcal{C}} f(x) := g(x) - h(x), \tag{1}$$

where $\mathcal{C} \subset \mathbb{R}^n$ is a compact and convex set, $g : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function and $h : \mathbb{R}^n \to \mathbb{R}$ is convex function possibly non-differentiable, with domain on $\mathbb{R}^n$. *Throughout the paper we assume that the gradient* ∇g *is L-Lipschitz continuous on* $\mathcal{C} \subset \mathbb{R}^n$, i.e., there exists a Lipschitz constant $L > 0$ such that

(A) $\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$ for all $x, y \in \mathcal{C}$.

Since we are assuming that $\mathcal{C} \subset \mathbb{R}^n$ is a compact set, its *diameter* is a finite number defined by

$$\operatorname{diam}(\mathcal{C}) := \max\{\|x - y\| : \ x, y \in \mathcal{C}\}.$$

Since $\mathcal{C} \subset \mathbb{R}^n$ is a compact, the study of problem (1) is bounded from below. Then, optimum value of the problem (1) satisfy $+\infty < f^* := \inf_{x \in \mathcal{C}} f(x)$ and optimal set $\mathcal{C}^*$ is non-empty. According to Theorem 2.3, the *first-order optimality condition* for problem (1) is stated as

$$(\nabla g(\bar{x}) - \bar{u})^T (x - \bar{x}) \geq 0, \qquad \forall \bar{u} \in \partial g(\bar{x}), \quad \forall x \in \mathcal{C}. \tag{2}$$

In general, the condition (2) is necessary but not sufficient for optimality. A point $\bar{x} \in \mathcal{C}$ satisfying condition (2) is called a *stationary point* to problem (1). Consequently, all $x^* \in \mathcal{C}^*$ satisfies (2). We finish this section by presenting a variant of a classical example of DC optimization problem that fits the aforementioned requirements.

Example 3.1. Let $\mathcal{C}_i \subset \mathbb{R}^n$ be an arbitrary set, for all $i = 1, \ldots, m$. The *square distance with respect to the set* $\mathcal{C}_i$, denoted by $d^2_{C_i} : \mathbb{R}^n \to \mathbb{R}$ is defined

$$d^2_{C_i}(x) := \inf_{y \in C_i} \|x - y\|^2. \tag{3}$$

En general, the function $d^2_{C_i}$ is not convex. However, due to $\|x - y\|^2 = \|x\|^2 - 2x^T y + \|y\|^2$, the distance function $d^2_{C_i}$ becomes to

$$d^2_{C_i}(x) := \|x\|^2 - \sup_{y \in C_i} \left(2x^T y - \|y\|^2\right), \tag{4}$$

which is a difference of two convex functions, see [22, Example 2.1.4]. Indeed, $d^2_{C_i} = g - h$, where $g(x) = \|x\|^2$ is a convex quadratic function and $h(x) = \sup_{y \in C_i} \left(2x^T y - \|y\|^2\right)$ is a convex function given by the supremum of affine functions $x \mapsto \ell_y(x) := 2x^T y - \|y\|^2$, for $y \in C_i$. Let us state the *constrained generalized Fermat-Weber location problem with* *square distances*. For that, take $\omega_i \geq 0$, $i = 1, \ldots, m$, such that $\sum_{i=1}^m \omega_i = 1$ and note that

$$\sum_{i=1}^m \alpha_i d_{C_i}^2(x) = \|x\|^2 - \sum_{i=1}^m \omega_i \sup_{y \in C_i} \left(2x^T y - \|y\|^2\right).$$

Thus, for a given constrained set compact and convex $\mathcal{C} \subset \mathbb{R}^n$, the DC version of the constrained generalized Fermat-Weber location problem is stated as follows

$$\min_{x \in \mathcal{C}} \left(\|x\|^2 - \textstyle\sum_{i=1}^m \omega_i \sup_{y \in C_i} \left(2x^T y - \|y\|^2\right)\right). \tag{5}$$

Finally, note the objective function of the last problem satisfies all conditions in problem 1.

4 Classical Frank–Wolfe algorithm with an adaptive stepsize

In this section, we formulate the *classic Frank-Wolfe algorithm* to solve problem (1) with an adaptive stepsize and provide a convergence analysis. Although we assume condition **(A)** for g, we will not use the Lipschitz constant to compute the stepsize in the formulation of the algorithm. The stepsize will be computed adaptively, by using a scheme introduced in [2, 4] that approximates the Lipschitz constant. The analysis of convergence of the method presented here shows that the convergence rate for weak-star-convex functions, despite not being convex, is still $\mathcal{O}(1/k)$ for both function values and duality gap. It should be mentioned that a version of the classic Frank-Wolfe algorithm has been proposed to address the problem (1) in [24]. The algorithm proposed incorporates a step size that depends on the curvature/Lipschitz-type constant of the gradient of $f = g - h$. According to the convergence analysis for non-convex objective functions (eg, [26]), it has been shown that the algorithm produces a stationary point with a convergence rate of only $\mathcal{O}(1/\sqrt{k})$. For stating the algorithm, we should assume that there exists a linear optimization oracle (LO oracle) capable of minimizing linear functions over the set $\mathcal{C}$. The statement of the algorithm is as follows:

Algorithm 1. **Frank–Wolfe**$_{C,f:=g-h}$ **algorithm**

Step 0. Select $x^0 \in \mathcal{C}$ and $L_0 > 0$. Set $k = 0$.

Step 1. Take $u^k \in \partial h(x^k)$. Set $j := \min\{\ell \in \mathbb{N} : \; 2^\ell L_k \geq 2L_0\}$.

Step 2. Use an "LO oracle" to compute an optimal solution p^k and the optimal value $\omega(x_k)$ as follows

$$p^k \in \operatorname{argmin}_{p \in \mathcal{C}} (\nabla g(x^k) - u^k)^{\mathrm{T}}(p - x^k), \qquad \omega(x_k) := (\nabla g(x^k) - u^k)^{\mathrm{T}}(p^k - x^k). \tag{6}$$

Step 3. If $\omega(x_k) = 0$, then **stop**. Otherwise, compute the step size $\lambda_j \in (0, 1]$ as follows

$$\lambda_j = \min\left\{1, \frac{|\omega(x_k)|}{2^j L_k \|p^k - x^k\|^2}\right\} := \operatorname{argmin}_{\lambda \in (0,1]} \left\{-|\omega(x_k)|\lambda + \frac{2^j L_k}{2}\|p^k - x^k\|^2 \lambda^2\right\}. \tag{7}$$

Step 4. If

$$f(x^k + \lambda_j(p^k - x^k)) \leq f(x^k) - |\omega(x_k)|\lambda_j + \frac{2^j L_k}{2}\|p^k - x^k\|^2 \lambda_j^2, \tag{8}$$

then set $j_k = j$ and go to **Step 5**. Otherwise, set $j = j + 1$ and go to **Step 3**.

Step 5. Set $\lambda_k := \lambda_{j_k}$ and define the next iterate x^{k+1} and the next approximation to the Lipschitz constant L_{k+1} as follows

$$x^{k+1} := x^k + \lambda_k(p^k - x^k), \qquad L_{k+1} := 2^{j_k-1}L_k. \tag{9}$$

Set $k \leftarrow k+1$, and go to **Step 1**.

Since the convergence analysis of Algorithm 1 is similar to that presented in [3], we will not include the proofs of the results here. We will only include the convergence rate analysis for weak-star-convex functions, which will be introduced in the next section. In order to simplify the notations, from now on we will use the following notation:

$$\omega_k := \omega(x_k). \tag{10}$$

Since $\omega_k = 0$ implies that x^k is a stationary point to problem (1). Thus, *in view of* (6) *from now on we assume that* $\omega_k < 0$*, for all* $k \in \mathbb{N}$. Next result establishes that the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 is well defined.

Proposition 4.1. *For all* $j \in \mathbb{N}$ *such that* $2^j L_k \geq L$*, the inequality* (8) *holds. Consequently, the number* j_k *in* **Step 4** *is well defined. Furthermore,* j_k *is the smallest non-negative integer satisfying the following two conditions*

$$2^{j_k}L_k \geq 2L_0,$$

$$f(x^k + \lambda_k(p^k - x^k)) \leq f(x^k) - |\omega_k|\lambda_k + \frac{2^{j_k}L_k}{2}\|p^k - x^k\|^2\lambda_k^2. \tag{11}$$

In addition, the sequence $(x^k)_{k\in\mathbb{N}}$ *generated by Algorithm 1 is well defined. And the following inequality holds*

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}|\omega_k|\lambda_k, \qquad k = 0, 1, \ldots.$$

For simplifying the notations we define the following constants

$$\alpha := 2(L + L_0)\,\mathrm{diam}(\mathcal{C})^2 > 0.$$

Lemma 4.2. *The sequence* $(L_k)_{k\in\mathbb{N}^*}$ *satisfies the following inequalities*

$$L_0 \leq L_k \leq L + L_0, \qquad k = 0, 1, \ldots.$$

and the step size sequence $(\lambda_k)_{k\in\mathbb{N}}$ *satisfies*

$$\lambda_k \geq \min\left\{1, \frac{|\omega_k|}{\alpha}\right\}, \qquad k = 0, 1, \ldots.$$

Theorem 4.3. $\lim_{k\to+\infty}\omega_k = 0$*. Consequently, every limit point of the sequence* $(x^k)_{k\in\mathbb{N}}$ *generated by Algorithm 1 is a stationary point to problem* (1)*.*

Next, we present some iteration-complexity bounds for the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 1.

Theorem 4.4. *For every* $N \in \mathbb{N}$*, there holds*

$$\min\{|\omega_k| : \; k = 0, 1, \ldots, N\} \leq \max\left\{2(f(x^0) - f^*), \sqrt{2\alpha(f(x^0) - f^*)}\right\}\frac{1}{\sqrt{N+1}}.$$

Since $\mathcal{C}$ is compact and the functions g and h are convex, we can define the following constants:

$$\Theta := \min\left\{\frac{2}{\bar{\Gamma}\operatorname{diam}(\mathcal{C})}, \frac{1}{\alpha}\right\} \qquad \bar{\Gamma} := \max_{x\in\mathcal{C}}\{\|\nabla g(x)\| + \|u\| : \ u \in \partial h(x)\}, \qquad \beta := \frac{\bar{\Gamma}^2}{\Theta L_0}.$$

Theorem 4.5. *Let $\epsilon > 0$. Then, Algorithm 1 generates a point x^k such that $|\omega_k| \leq \epsilon$, performing at most,*

$$\left(\max\left\{\Gamma^2, \alpha\Gamma\right\}\frac{1}{\epsilon^2}\right)\left(2 + \log_2\left(\beta\frac{1}{\epsilon^2}\right)\right) = \mathcal{O}(|\log_2(\epsilon)|\epsilon^{-2})$$

evaluations of functions f, and $\max\left\{\Gamma^2, \alpha\Gamma\right\}\frac{1}{\epsilon^2} + 1 = \mathcal{O}(\epsilon^{-2})$ evaluations of gradients of f and subgradients of h.

4.1 Iteration-complexity analysis: The weak-star-convex case

In this section we present iteration-complexity bounds for the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 when the objective function $f = g - h$ is a weak-star-convex function and has Lipschitz continuous gradient in $\mathcal{C} \subset \mathbb{R}^n$. Let us begin by recalling the concept of the star-convex function introduced in [34].

Definition 4.6. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be star-convex in $\mathcal{C}$ if its set of global minima X^* on the set $\mathcal{C}$ is not empty and for any $x^* \in X^*$ we have

$$f(\lambda x^* + (1-\lambda)x) \leq \lambda f(x^*) + (1-\lambda)f(x), \qquad \forall x \in \mathcal{C}, \forall\lambda \in [0,1]. \tag{12}$$

Every convex function with global minimizer set non-empty is a star-convex function, but in general, star-convex functions need not be convex. In the following we present two examples of star-convex functions that are not convex, which appeared in [34]. We show that these examples are difference of two convex functions. Although the first example does not fit the assumptions of problem 1 we include it here because it is a classical example of star-convex function.

Example 4.7. The star-convex function $\phi(t) = |t|(1 - e^{-|t|})$ is not convex. In addition, ϕ is a difference of two convex functions. Indeed, considering the following two convex functions $\varphi(t) = |t|(1 - e^{-|t|}) + e^{|t|}$ and $\psi(t) = e^{|t|}$, we have $\phi = \varphi - \psi$.

Example 4.8. Consider the star-convex function $f(s,t) = s^2t^2 + s^2 + t^2$. The function f is not convex, but is a difference of two convex functions. Indeed, letting the following convex functions $g(s,t) = (s^2+t^2)^2+s^2+t^2$ and $h(s,t) = s^2t^2+s^4+t^4$ we have $f = g-h$.

Let us introduce the weak-star-convex function notion, which generalizes the star-convex function concept proposed in [34] as well as a few other related concepts, such as [21, 42].

Definition 4.9. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be weak-star-convex in $\mathcal{C}$ if its set of global minima X^* on the set $\mathcal{C}$ is not empty and, for each $x \in \mathcal{C}$, there exists $x^*_x \in X^*$ such that

$$f(\lambda x^*_x + (1-\lambda)x) \leq \lambda f(x^*_x) + (1-\lambda)f(x), \qquad \forall\lambda \in [0,1]. \tag{13}$$

As aforementioned, star-convex functions are not needed to be convex, but it should be noted that every star-convex function is a weak star-convex function. In fact, the minimizer x_x^* used in Definition 4.9 depends on the point x. We outline a general procedure for generating star-convex functions in the next examples.

Example 4.10. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is not necessarily star-convex in $\mathcal{C}$. The collection of global minima of f on the set $\mathcal{C}$ is denoted by the set X^*. Assume that the following conditions holds:

(i) the set X^* is not empty, i.e., $X^* \neq \varnothing$;

(ii) $\mathcal{C} := \cup_{i=1}^n \Omega_i$, where Ω_i is a convex set, for all $i = 1, \ldots, n$;

(iii) there exists $x_1^*, \ldots, x_n^* \in X^*$ such that $x_i^* \in \Omega_i$, for all $i = 1, \ldots, n$;

(iv) the function f is convex on Ω_i, for all $i = 1, \ldots, n$.

Under the condition $(i), (ii), (iii)$ and (iv) we can show that the function f is weak-star-convex. Indeed, take $x \in \mathcal{C}$. The items (ii) implies that the exists i such that $x \in \Omega_i$. Moreover, the items (i) and (ii) implies that there exists a minimizer of f such that $x_i^* \in \Omega_i$. Since f is convex in the convex set Ω_i the inequality (13) holds for $x_x^* = x_i^*$. Therefore, f is weak-star-convex. For instance, let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$\varphi(t, s) = \frac{1}{2}\left(t^2 + s^2\right) - |t| - |s|. \tag{14}$$

The function φ is weak-star-convex but not star-convex. In fact, the minimizer set of φ is given by $X^* = \{x_1^* := (1, 1), x_2^* := (-1, 1), x_3^* := (1, -1), x_4^* := (-1, -1)\}$, which satisfies the item (i). Set $\Omega_1 := \{(s, t) \in \mathbb{R}^2 : \ s \geq 0, t \geq 0\}$, $\Omega_2 := \{(s, t) \in \mathbb{R}^2 : \ -s \geq 0, t \geq 0\}$, $\Omega_3 := \{(s, t) \in \mathbb{R}^2 : \ s \geq 0, -t \geq 0\}$ and $\Omega_4 := \{(s, t) \in \mathbb{R}^2 : \ -s \geq 0, -t \geq 0\}$. Since the sets Ω_1, Ω_2, Ω_3 and Ω_4 are convex and $\mathbb{R}^2 := \cup_{i=1}^4 \Omega_i$, the item (ii) is satisfied. We also have $x_i^* \in \Omega_i$, for $i = 1, 2, 3, 4$. Consequently, the item (iii) is also satisfied. Finally, note that $\varphi(t, s) = (1/2)\left(t^2 + s^2\right) - t - s$, for all $(s, t) \in \Omega_1$, $\varphi(t, s) = (1/2)\left(t^2 + s^2\right) + t - s$, for all $(s, t) \in \Omega_2$, $\varphi(t, s) = (1/2)\left(t^2 + s^2\right) - t + s$, for all $(s, t) \in \Omega_2$ and $\varphi(t, s) = (1/2)\left(t^2 + s^2\right) + t + s$, for all $(s, t) \in \Omega_1$. Hence, we obtain that f is convex on the set Ω_i, for $i = 1, 2, 3, 4$. Thus, the item (iv) is satisfied. Therefore, because items (i),(ii), (iii) and (iv) are satisfied, we conclude that the function (14) is weak-star-convex. Since

$$0 = f\Big(-\frac{1}{2} + \frac{1}{2}\Big) > \frac{1}{2}f(-1) + \frac{1}{2}f(1) = -1,$$

we conclude that f is not star-convex. In general, with a little more effort, we can show that the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \alpha\|x\|^2 - \beta\|x\|_1$, where $\alpha > 0$ and $\beta > 0$, is weak-star-convex but not star-convex.

Example 4.11. In some cases the generalised Fermat-Weber problem (5) is weak star convex. For example, when $m = 1$ and C_1 is a finite union of compact convex sets such that every set in the union contains a solution from X^*, then the objective is weak-star-convex.

Proposition 4.12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. Assume that the set of global minimum of f on the set $\mathcal{C}$, denoted by the set X^*, is non-empty, and let f^* be the minimum value of f on the set $\mathcal{C}$. If f is weak-star-convex in $\mathcal{C}$, then for each $x \in \mathcal{C}$ there exists $x^*_x \in X^*$ such that $f^* - f(x) \geq v^{\mathrm{T}}(x^*_x - x)$, for all $v \in \partial_c f(x)$.*

Proof. Let $x \in \mathcal{C}$. It follows from Definition 4.9 that there exists a minimizer $x^*_x \in X^*$ of f such that $f(x^*_x) - f(x) \geq (f(x + \lambda(x^*_x - x)) - f(x))/\lambda$. Then, taking the limit as λ goes to $+0$ we have $f(x^*_x) - f(x) \geq f^\circ(x; x^*_x - x)$. Since $f^* = f(x^*)$, by using the last statement in Theorem 2.1 yields the desired inequality. □

Corollary 4.13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. If f is weak-star-convex in $\mathcal{C}$, then each stationary point $\bar{x} \in \mathcal{C}$ of the problem $\min_{x \in \mathcal{C}} f(x)$ is also a global minimizer.*

Proof. Let $\bar{x} \in \mathcal{C}$ be a stationary point of the problem $\min_{x\in\mathcal{C}} f(x)$, denotes by X^* the set of its global solution and by f^* the global minimum value. Thus, we have

$$v^T(x - \bar{x}) \geq 0, \qquad \forall v \in \partial_c f(\bar{x}), \qquad \forall\, x \in \mathcal{C}. \tag{15}$$

Since f is weak-star-convex in $\mathcal{C}$, it follows from Proposition 4.12 and Theorem 2.2 that there exists $x^*_{\bar{x}} \in \mathcal{C}^*$ such that $f^* - f(\bar{x}) \geq f^\circ(\bar{x}; x^*_{\bar{x}} - \bar{x}) \geq \bar{v}^T(x^*_{\bar{x}} - \bar{x})$. Thus, by using (15) we conclude that $f^* \geq f(\bar{x})$. Considering that f^* the global minimum value, we have $f(\bar{x}) = f^*$. Therefore, $\bar{x}$ is also a global minimizer. □

Theorem 4.14. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is weak-star-convex in $\mathcal{C}$. Let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then,*

(i) $f(x^k) - f^* \leq \dfrac{4(L + L_0)\,\mathrm{diam}(\mathcal{C})^2}{k}$, *for all* $k = 1, 2, \ldots$.

(ii) $\displaystyle\min_{\ell \in \{\lfloor \frac{k}{2} \rfloor + 2, \ldots, k\}} |\omega_\ell| \leq \frac{16(L + L_0)\,\mathrm{diam}(\mathcal{C})^2}{k - 2}$, *for all* $k = 3, 4, \ldots$, *where* $\lfloor k/2 \rfloor = \max_{n\in\mathbb{N}} \{n \leq k/2\}$.

Proof. It follows from (11) in Proposition 4.1 that

$$f(x^k + \lambda_k(p^k - x^k)) \leq f(x^k) - |\omega(x_k)|\lambda_k + \frac{2^{j_k} L_k}{2}\|p^k - x^k\|^2\lambda_k^2. \tag{16}$$

On the other hand, by using (7) we conclude that

$$\lambda_k = \mathrm{argmin}_{\lambda\in(0,1]} \left\{ -|\omega(x_k)|\lambda + \frac{2^{j_k} L_k}{2}\|p^k - x^k\|^2\lambda^2 \right\}.$$

Hence, letting $\beta_k := 2/(k + 2)$, it follows from (16) and the last inequality that

$$f(x^k + \lambda_k(p^k - x^k)) \leq f(x^k) - |\omega_k|\beta_k + \frac{2^{j_k} L_k}{2}\|p^k - x^k\|^2\beta_k^2.$$

Since $\|p(x^k) - x^k\| \leq \mathrm{diam}(\mathcal{C})$, the last inequality together with (9) and Lemma (4.2) yield

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* - |\omega_k|\beta_k + (L + L_0)\,\mathrm{diam}(\mathcal{C})^2\beta_k^2.$$

Taking into account that f is a weak-star-convex function in $\mathcal{C}$, it follows from Proposition 4.12 and Theorem 2.2 that there exists $x_k^* \in \mathcal{C}^*$ such that

$$f^* - f(x^k) \geq (\nabla g(x^k) - u^k)^{\mathrm{T}}(x_k^* - x^k).$$

Thus, it follows from (6) and (10) that $0 \geq f^* - f(x^k) \geq \omega_k$, which implies that $0 \leq f(x^k) - f^* \leq |\omega_k|$. Thus, applying Lemma 2.8 with

$$a_k := f(x^k) - f^* \leq b_k := |\omega_k|, \qquad A := 2(L + L_0)\,\mathrm{diam}(\mathcal{C})^2,$$

the desired inequalities follows. $\square$

According to Theorem 4.14, functions with the weak-star-covexity property allow the Frank-Wolfe method to efficiently minimize them even when their landscape is not convex.

5 Frank–Wolfe algorithm with finite difference

In this section, based on Algorithm 1, we formulate a *version of Frank–Wolfe algorithm to solve problem* (1) *with finite difference approximating the gradient* ∇g *and with an adaptive step size.* For that, let us first recall de concept of finite difference of a continuously differentiable $g : \mathbb{R}^n \to \mathbb{R}$. Let $D(x, h)$ be the *vector finite difference of* g *at* x *with increment* $s > 0$, which approximates the gradient $\nabla g(x)$, defined by

$$D(x,s) := (D_1(x,s), \ldots, D_n(x,s)), \qquad D_i(x,s) := \frac{g(x + se_i) - g(x)}{s}, \quad i = 1, \ldots, n. \tag{17}$$

The next lemma gives us an upper bound for the approximation error of the gradient of the function ∇g by its finite difference $D(x, h)$, see for example [35, (8.4) p. 195].

Lemma 5.1. *Assume that* g *satisfies the assumption* **(A)**. *Then, there holds*

$$\|\nabla g(x) - D(x,s)\| \leq \frac{\sqrt{n}L}{2}s, \qquad \forall x \in \mathbb{R}^n, \quad \forall s > 0.$$

Proof. Setting $\nabla g(x) = (\partial_1 g(x), \ldots, \partial_n g(x))$ and using (17) we have

$$\|\nabla g(x) - D(x,s)\|^2 \leq \sum_{i=1}^{n} \left(\partial_i g(x) - D_i(x,s)\right)^2. \tag{18}$$

Applying Proposition 2.4 with $v = e_i$ and $\lambda = s$, we conclude that $D_i(x,s) - \partial_i g(x) \leq (Ls)/2$. Thus, (18) implies that $\|\nabla g(x) - D(x,s)\|^2 \leq n((Ls)/2)^2$, which implies the desired inequality. $\square$

The statement of the Frank–Wolfe algorithm to solve problem (1) with finite difference approximating the gradient ∇g is as follows:

Algorithm 2. **Frank–Wolfe**$_{C,f:=g-h}$ **algorithm with finite difference**

Step 0. Select $x^0, x^1 \in \mathcal{C}$, $x^0 \neq x^1$ and $L_1 > 0$. Set $k = 1$.

Step 1. Take $u^k \in \partial h(x^k)$. Set $j := \min\{\ell \in \mathbb{N} :\ 2^\ell L_k \geq 2L_1\}$ and $i = 0$.

Step 2. Set
$$s_{ij} = \frac{2L_1}{2^{i+j}L_k\sqrt{n}}\|x^k - x^{k-1}\|. \tag{19}$$

Step 3. Use an "LO oracle" to compute an optimal solution p^{ij} and the optimal value $\omega_{ij}(x_k)$ as follows
$$p^{ij} \in \mathrm{argmin}_{p\in\mathcal{C}}(D(x^k, s_{ij}) - u^k)^{\mathrm{T}}(p - x^k), \qquad \omega_{ij}(x_k) := (D(x^k, s_{ij}) - u^k)^{\mathrm{T}}(p^{ij} - x^k). \tag{20}$$

Step 4. If $\omega_{ij}(x_k) < 0$, then set $i_k = i$ and go to **Step 5**. Otherwise, set $i \leftarrow i+1$, and go to **Step 2**.

Step 5. Compute the step size $\lambda_j \in (0, 1]$ as follows
$$\lambda_j = \min\left\{1, \frac{|\omega_{i_kj}(x_k)|}{2^jL_k\|p^{i_kj} - x^k\|^2}\right\} := \mathrm{argmin}_{\lambda\in(0,1]}\left\{-|\omega_{i_kj}(x_k)|\lambda + \frac{2^jL_k}{2}\|p^{i_kj} - x^k\|^2\lambda^2\right\}. \tag{21}$$

Step 6. If
$$f(x^k + \lambda_j(p^{i_kj} - x^k)) \le f(x^k) - \frac{1}{2}|\omega_{i_kj}(x_k)|\lambda_j - \frac{2^jL_k}{4}\|p^{i_kj} - x^k\|^2\lambda_j^2 + \frac{L_1}{4}\|x^k - x^{k-1}\|^2, \tag{22}$$
then set $j_k = j$ and go to **Step 7**. Otherwise, set $j \leftarrow j+1$ and go to **Step 2**.

Step 7. Set $\lambda_k := \lambda_{j_k}$, $p^k := p^{i_kj_k}$, define the next iterate x^{k+1} and the next approximation to the Lipschitz constant L_{k+1} as follows
$$x^{k+1} := x^k + \lambda_k(p^k - x^k), \qquad L_{k+1} := 2^{j_k-1}L_k. \tag{23}$$
Set $k \leftarrow k+1$, and go to **Step 1**.

Let us describe the main features of Algorithm 2. First of all, note that due to $x^0, x^1 \in \mathcal{C}$, $p^k \in \mathcal{C}$ and $\lambda_k \in (0, 1]$ for all $k = 0, 1, \ldots$, and $\mathcal{C}$ be convex set, it follows from inductive arguments and first equality in (23) that $(x^k)_{k\in\mathbb{R}} \subset \mathcal{C}$. Note that (20) implies that
$$\omega_{ij}(x_k) \le D(x^k, s_{ij})^T(p - x^k), \qquad p \in \mathcal{C}. \tag{24}$$
Hence, considering that $x^k \in \mathcal{C}$, letting $p = x^k$ in the last inequality we conclude that
$$\omega_{ij}(x_k) \le 0. \tag{25}$$
In order to simplify the notations, from now on we will use the following two notations:
$$s_k := s_{i_kj_k}, \qquad \omega_k := \omega_{i_kj_k}(x_k). \tag{26}$$

Proposition 5.2. *Assume that in* **Step 3** *we have $\omega_{ij}(x_k) = 0$, for all $i \in \mathbb{N}$. Then,*
$$(\nabla g(x^k) - u^k)^{\mathrm{T}}(p - x^k) \ge 0, \qquad \forall p \in \mathcal{C}, \tag{27}$$
i.e., x^k is a stationary point to problem (1).

Proof. If in **Step 3** we have $\omega_{ij}(x_k) = 0$, for all $i \in \mathbb{N}$, then it follows from (24) that
$$0 \le (D(x^k, s_{ij}) - u^k)^{\mathrm{T}}(p - x^k), \qquad \forall i \in \mathbb{N}, \quad \forall p \in \mathcal{C}. \tag{28}$$
Since (19) implies that $\lim_{i\to+\infty} s_{ij} = 0$, it follows from (17) that $\lim_{i\to+\infty} D(x^k, s_{ij}) = \nabla g(x^k)$. Hence, taking limit in (28) we conclude that (27) holds. □

In view of Proposition 5.2, (25) *and* (26) *from now on we assume that $\omega_k < 0$, for all $k \in \mathbb{N}$.*

5.1 Well-definedness

Herein, we establish that the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 is well defined.

Lemma 5.3. *Let s_{ij} be define in* (19) *and assume that $2^j L_k \geq 2L$. Then, there holds*

$$\|\nabla g(x^k) - D(x^k, s_{ij})\| \leq \frac{L_1}{2}\|x^k - x^{k-1}\|.$$

Proof. By applying Lemma 5.1 with $x = x^k$ and $s = s_{ij}$we obtain that

$$\|\nabla g(x^k) - D(x^k, s_{ij})\| \leq \frac{\sqrt{n}L}{2} s_{ij}$$

Therefore, using (19) and considering that $2^j L_k \geq 2L$ the desired inequality follows. □

Lemma 5.4. *Let $j \in \mathbb{N}$ such that $2^j L_k \geq 2(L + L_1/2)$. Then,* (22) *holds and j_k is the smallest $j \in \mathbb{N}$ satisfying the following two conditions $2^j L_k \geq 2L_1$ and* (22)*. Consequently, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 is well defined and satisfies the following inequality:*

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}|\omega_k|\lambda_k - \frac{L_{k+1}}{2}\|x^{k+1} - x^k\|^2 + \frac{L_1}{4}\|x^k - x^{k-1}\|^2, \qquad k = 1, 2, \ldots. \tag{29}$$

Proof. It follows from Proposition 2.4 that

$$g(x^k + \lambda_j(p^{i_k j} - x^k)) \leq g(x^k) + \nabla g(x^k)^{\mathrm{T}}(p^{i_k j} - x^k)\lambda_j + \frac{L}{2}\|p^{i_k j} - x^k\|^2\lambda_j^2. \tag{30}$$

On the other hand, considering that $u^k \in \partial h(x^k)$ it follows from Proposition (2.5) that

$$h(x^k + \lambda_j(p^{i_k j} - x^k)) \geq h(x^k) + \lambda_j\langle u^k, p^{i_k j} - x^k\rangle \tag{31}$$

Thus, combining (30) with (31) and taking into account the equality in (20) and (25), some algebraic manipulations shows that

$$\begin{aligned} f(x^k + \lambda_j(p^{i_k j} - x^k)) \leq f(x^k) - |\omega_{i_k j}(x_k)|\lambda_j + \frac{2^j L_k}{2}\|p^{i_k j} - x^k\|^2\lambda_j^2 + \\ (\nabla g(x^k) - D(x^k, s_{i_k j}))^{\mathrm{T}}(p^{i_k j} - x^k)\lambda_j + \frac{L - 2^j L_k}{2}\|p^{i_k j} - x^k\|^2\lambda_j^2. \end{aligned} \tag{32}$$

To proceed with the proof we need first to prove that for λ_j defined in (21) we have

$$-|\omega_{i_k j}(x_k)|\lambda_j + \frac{2^j L_k}{2}\|p^{i_k j} - x^k\|^2\lambda_j^2 \leq -\frac{1}{2}|\omega_{i_k j}(x_k)|\lambda_j. \tag{33}$$

For that we separately consider two cases: $\lambda_j = |\omega_{i_k j}(x_k)|/(2^j L_k\|p^{i_k j} - x^k\|^2)$ and $\lambda_j = 1$. In the former, we have $2^j L_k\|p^{i_k j} - x^k\|^2 = |\omega_{i_k j}(x_k)|/\lambda_j$,which substituting into the left hand side of (33) yields an equality. If now $\lambda_j = 1$, then it follows from (21) that $\lambda_j = 1 \leq |\omega_{i_k j}(x_k)|/(2^j L_k\|p^{i_k j} - x^k\|^2)$ or equivalently $2^j L_k\|p^{i_k j} - x^k\|^2 \leq |\omega_{i_k j}(x_k)|$ and

$\lambda_j = 1$, which again substituting into the left hand side of (33) yields the inequality. Hence, (33) holds. Therefore, (32) becomes

$$f(x^k + \lambda_j(p^{i_k j} - x^k)) \leq f(x^k) - \frac{1}{2}|\omega_{i_k j}(x_k)|\lambda_j + \\ (\nabla g(x^k) - D(x^k, s_{i_k j}))^{\mathrm{T}}(p^{i_k j} - x^k)\lambda_j + \frac{L - 2^j L_k}{2}\|p^{i_k j} - x^k\|^2\lambda_j^2. \quad (34)$$

Since $2^j L_k \geq 2(L + L_1/2) > L$, we can apply Lemma 5.3 to obtain that

$$(\nabla g(x^k) - D(x^k, s_{i_k j}))^{\mathrm{T}}(p^{i_k j} - x^k)\lambda_j \leq \frac{L_1}{2}\|x^k - x^{k-1}\|\|p^{i_k j} - x^k\|\lambda_j.$$

Hence, considering that $2\|x^k - x^{k-1}\|\|p^{i_k j} - x^k\|\lambda_j \leq \|x^k - x^{k-1}\|^2 + \|p^{i_k j} - x^k\|^2\lambda_j^2$, we have

$$(\nabla g(x^k) - D(x^k, s_{i_k j}))^{\mathrm{T}}(p^{i_k j} - x^k)\lambda_j \leq \frac{L_1}{4}\|p^{i_k j} - x^k\|^2\lambda_j^2 + \frac{L_1}{4}\|x^k - x^{k-1}\|^2.$$

Thus, combining the last inequality with (34) and considering that $\lambda_j < 1$, we conclude that

$$f(x^k+\lambda_j(p^{i_k j}-x^k)) \leq f(x^k)-\frac{1}{2}|\omega_{i_k j}(x_k)|\lambda_j+\frac{L + L_1/2 - 2^j L_k}{2}\|p^{i_k j}-x^k\|^2\lambda_j^2+\frac{L_1}{4}\|x^k-x^{k-1}\|^2.$$

Considering that $2^j L_k \geq 2(L + L_1/2)$, the last inequality implies that (22) holds. Since the first trial j in (22) is defined in **Step 1**, i.e., $j := \min\{\ell \in \mathbb{N} : \ 2^\ell L_k \geq 2L_1\}$ and $2^{j+1}L_k \geq 2^j L_k$ we conclude that j_k is the smallest $j \in \mathbb{N}$ satisfying $2^j L_k \geq 2L_1$ and (22). Consequently, $(x^k)_{k\in\mathbb{N}}$ is well defined, which proof the first three statements of the lemma. Therefore, (29) follows from the definition of j_k in **Step 6**, (22) and (23), which concludes the proof. □

Since the proof of the next lemma is similar to Lemma 4.2 it will be omitted.

Lemma 5.5. *The sequence $(L_k)_{k\in\mathbb{N}^*}$ satisfies the following inequalities*

$$L_1 \leq L_k \leq 2(L + L_1), \qquad k = 1, 2, \ldots. \quad (35)$$

and the step size sequence $(\lambda_k)_{k\in\mathbb{N}^*}$ satisfies

$$\lambda_k \geq \min\left\{1, \frac{|\omega_k|}{2(L + L_1)\,\mathrm{diam}(\mathcal{C})^2}\right\}, \qquad k = 1, 2, \ldots. \quad (36)$$

5.2 Convergence analysis

In this section we analyze the convergence properties of the sequence $(x^k)_{k\in\mathbb{N}}$. For that, we assume that the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 is infinite.

Lemma 5.6. *The sequence $\big(f(x^k) + (L_1/2)\|x^k - x^{k-1}\|^2\big)_{k\in\mathbb{N}}$ is monotone and non-increasing. Moreover, there holds:*

$$\lim_{k\to+\infty} \|x^k - x^{k-1}\| = 0. \quad (37)$$

As a consequence, $\lim_{k\to+\infty} f(x^k) = \bar{f}$ for some $\bar{f} \geq f^*$.

Proof. Since $L_k \geq L_1$, for all $k \in \mathbb{N}$, it follows from Lemma 5.4 that

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}|\omega_k|\lambda_k - \frac{L_1}{2}\|x^{k+1} - x^k\|^2 + \frac{L_1}{2}\|x^k - x^{k-1}\|^2, \qquad k = 1, 2, \ldots.$$

which, taking into account that $\frac{1}{2}|\omega_k|\lambda_k \geq 0$, yields

$$f(x^{k+1}) + \frac{L_1}{2}\|x^{k+1} - x^k\|^2 \leq f(x^k) + \frac{L_1}{2}\|x^k - x^{k-1}\|^2, \qquad k = 1, 2, \ldots..$$

which implies the first statement of the lemma. We proceed to prove the second statement. For that we first note that Lemma 5.5 implies that $L_1 \leq L_k$, for all $k \in \mathbb{N}$. Thus, using again that $\frac{1}{2}|\omega_k|\lambda_k \geq 0$, it follows from (29) that

$$\frac{L_{k+1}}{2}\|x^{k+1} - x^k\|^2 - \frac{L_k}{4}\|x^k - x^{k-1}\|^2 \leq f(x^k) - f(x^{k+1}), \qquad k = 1, 2, \ldots. \tag{38}$$

Let $\ell \in \mathbb{N}$ be such that $\ell > 1$. Summing up the inequality (38) from 1 to ℓ we obtain

$$\sum_{k=1}^{\ell} \Big(\frac{L_{k+1}}{2}\|x^{k+1} - x^k\|^2 - \frac{L_k}{4}\|x^k - x^{k-1}\|^2\Big) \leq f(x^1) - f^*. \tag{39}$$

Considering that $L_1 \leq L_k$, for all $k \in \mathbb{N}$, some algebraic manipulations show that

$$-\frac{L_1}{4}\|x^1 - x^0\|^2 + \frac{L_1}{4}\sum_{k=1}^{\ell}\|x^{k+1} - x^k\|^2 \leq \sum_{k=1}^{\ell}\Big(\frac{L_{k+1}}{2}\|x^{k+1} - x^k\|^2 - \frac{L_k}{4}\|x^k - x^{k-1}\|^2\Big).$$

Combining (39) with the last inequality we obtain that

$$\sum_{k=1}^{\ell}\|x^{k+1} - x^k\|^2 \leq \frac{4}{L_1}\Big(f(x^1) - f^* + \frac{L_1}{4}\|x^1 - x^0\|^2\Big).$$

Since the last inequality holds for all $\ell \in \mathbb{N}$, we have $\sum_{k=1}^{+\infty}\|x^{k+1} - x^k\|^2 \leq +\infty$. Thus, (37) holds. To prove the last statement, we first note that due to $\big(f(x^k) + (L_1/2)\|x^k - x^{k-1}\|^2\big)_{k\in\mathbb{N}}$ be a non-increasing sequence and bounded from below by f^*, it must be convergent. Thus, we set $\bar{f} = \lim_{k\to+\infty}\big(f(x^k) + (L_1/2)\|x^k - x^{k-1}\|^2\big)$. Hence, considering that

$$f(x^k) = f(x^k) + (L_1/2)\|x^k - x^{k-1}\|^2 - (L_1/2)\|x^k - x^{k-1}\|^2 \geq f^*,$$

by using (37) we conclude that $\lim_{k\to+\infty} f(x^k) = \bar{f} \geq f^*$ and the proof is concluded. □

For simplifying the notations we define the following constants

$$\bar{\alpha} := 2(L + L_1)\,\mathrm{diam}(\mathcal{C})^2 > 0. \tag{40}$$

Lemma 5.7. *The following inequality holds:*

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}\min\Big\{|\omega_k|, \frac{1}{\bar{\alpha}}|\omega_k|^2\Big\} - \frac{L_1}{2}\|x^{k+1} - x^k\|^2 + \frac{L_1}{2}\|x^k - x^{k-1}\|^2, \tag{41}$$

for all $k = 1, 2, \ldots$.

Proof. Combining inequalities (29) with (35) and (36), and also using (40), we conclude that

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}|\omega_k| \min\left\{1, \frac{1}{\bar{\alpha}}|\omega_k|\right\} - \frac{L_1}{2}\|x^{k+1} - x^k\|^2 + \frac{L_1}{4}\|x^k - x^{k-1}\|^2,$$

for all $k = 1, 2, \ldots$. Thus, (41) is an immediate consequence of the previous inequality. □

In the next theorem we will use the following constant

$$\Gamma := 2(f(x^1) - f^*) + L_1\|x^2 - x^1\|^2. \tag{42}$$

Theorem 5.8. *The following statements hold:*

(i) $\lim_{k\to+\infty} \omega_k = 0$. *Consequently, every limit point of the sequence* $(x^k)_{k\in\mathbb{N}}$ *generated by Algorithm 2 is a stationary point to problem* (1)*;*

(ii) for every $N \in \mathbb{N}$, *there holds*

$$\min\{|\omega_k| : \ k = 1, \ldots, N\} \leq \max\left\{\Gamma, \sqrt{\bar{\alpha}\Gamma}\right\} \frac{1}{\sqrt{N}}.$$

Proof. To prove the item (i), we first note that (41) is equivalent to

$$\frac{1}{2}\min\left\{|\omega_k|, \frac{1}{\bar{\alpha}}|\omega_k|^2\right\} \leq \left(f(x^k)+\frac{L_1}{2}\|x^k-x^{k-1}\|^2\right)-\left(f(x^{k+1})+\frac{L_1}{2}\|x^{k+1}-x^k\|^2\right), \tag{43}$$

for all $k = 1, 2, \ldots$. It follows from Lemma 5.6 that $\left(f(x^k) + (L_1/2)\|x^k - x^{k-1}\|^2\right)_{k\in\mathbb{N}^*}$ monotone and non-increasing. Thus, considering that it is also bounded from below by f^*, we conclude that it converges. Hence, taking limit in (43) we have

$$\lim_{k\to+\infty} \min\left\{|\omega_k|, \frac{1}{\bar{\alpha}}|\omega_k|^2\right\} = 0,$$

which implies the first statement of item (i). For proving the second statement of item (i), take $\bar{x}$ a limit point of the sequence $(x^k)_{k\in\mathbb{N}}$ and subsequence $(x^{k_i})_{k\in\mathbb{N}}$ such that $\lim_{i\to+\infty} x^{k_i} = \bar{x}$. Since $\mathcal{C}$ is a compact set the sequence $(u^k)_{k\in\mathbb{N}^*}$ is bounded. Thus, without loss of generality we assume that $\lim_{i\to+\infty} u^{k_i} = \bar{u}$. On the other hand, Lemma 5.6 together with (19) imply that

$$\lim_{k\to+\infty} s_k = 0. \tag{44}$$

Let $(s_{k_i})_{i\in\mathbb{N}^*}$ subsequence of $(s_k)_{k\in\mathbb{N}^*}$. Thus, we conclude from (17) that

$$\lim_{i\to+\infty} D(x^{k_i}, s_{k_i}) = \nabla g(\bar{x}). \tag{45}$$

Finally, using (20) we conclude that

$$(D(x^{k_i}, s_{k_i}) - u^{k_i})^T(p - x^{k_i}) \geq \omega_{k_i}, \qquad \forall p \in \mathcal{C}.$$

Therefore, taking limite in the last inequality and using (44) together (45) we obtain that

$$(\nabla g(\bar{x}) - \bar{u})^T (p - \bar{x}) \geq 0, \qquad \bar{u} \in \partial g(\bar{x}), \qquad \forall\, x \in \mathcal{C},$$

which implies that $\bar{x}$ is a stationary point to problem (1). To prove the item (ii), take $N \in \mathbb{N}^*$. Thus, (43) implies that

$$\sum_{k=1}^{N} \frac{1}{2} \min\left\{ |\omega_k|, \frac{1}{\bar{\alpha}}|\omega_k|^2 \right\} \leq \left(f(x^1) + \frac{L_1}{2}\|x^2 - x^1\|^2 \right) - \left(f(x^{N+1}) + \frac{L_1}{2}\|x^N - x^{N-1}\|^2 \right).$$

Hence, we conclude that

$$\min\left\{ \frac{1}{2} \min\left\{ |\omega_k|, \frac{1}{\bar{\alpha}}|\omega_k|^2 \right\} :\; k = 1, \dots N \right\} \leq \frac{1}{N} \left(f(x^1) - f^* + \frac{L_1}{2}\|x^2 - x^1\|^2 \right).$$

It follows from the last inequality that there exists $\bar{k} \in \{1, \dots N\}$ such that

$$|\omega_{\bar{k}}| \leq \frac{1}{N}\left(2(f(x^1) - f^*) + L_1\|x^2 - x^1\|^2 \right), \qquad |\omega_{\bar{k}}| \leq \sqrt{\frac{\bar{\alpha}}{N}\left(2(f(x^1) - f^*) + L_1\|x^2 - x^1\|^2 \right)},$$

which, by using (42), implies the item (ii). □

6 Conclusions

We studied the asymptotic convergence properties and iteration complexity of Algorithm 1. In particular, for weak-star-convex objective functions, a rate of convergence is proven to be $\mathcal{O}(1/k)$ for both the function values and the duality gap. The study of Algorithm 2, on the other hand, became more complex because of the usage of finite difference, therefore we were only capable of providing its asymptotic convergence analysis, and rate of convergence for the duality gap is $\mathcal{O}(1/\sqrt{k})$. Then, an extensive iteration complexity study for this algorithm is absent, including weak-star-convex objective functions. In particular, there are thus various research opportunities to improve the current results for Algorithm 2.

Funding

The first and third authors were supported by the Australian Research Council (ARC), Solving hard Chebyshev approximation problems through nonsmooth analysis (Discovery Project DP180100602). The second author was supported in part by CNPq - Brazil Grants 304666/2021-1. This work was done, in part, while the second author visited the first and third authors at Deakin University in November 2022. The second author thanks the host institution for funding the visit and for the pleasant scientific atmosphere it provided during his visit.

References

[1] A. Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017.

[2] A. Beck, E. Pauwels, and S. Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM J. Optim.*, 25(4):2024–2049, 2015.

[3] A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Methods Oper. Res.*, 59(2):235–247, 2004.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[5] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, second edition, 1999.

[6] K. Boufi, M. El Haffari, and A. Roubi. Optimality conditions and a method of centers for minimax fractional programs with difference of convex functions. *J. Optim. Theory Appl.*, 187(1):105–132, 2020.

[7] A. Bouhamidi, M. Bellalij, R. Enkhbat, K. Jbilou, and M. Raydan. Conditional gradient method for double-convex fractional programming matrix problems. *J. Optim. Theory Appl.*, 176(1):163–177, 2018.

[8] N. Boyd, G. Schiebinger, and B. Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.*, 27(2):616–639, 2017.

[9] F. Clarke. *Optimization and Nonsmooth Analysis*. Canadian Mathematical Society series of monographs and advanced texts. Wiley, 1983.

[10] W. de Oliveira. Sequential difference-of-convex programming. *J. Optim. Theory Appl.*, 186(3):936–959, 2020.

[11] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Corrected reprint of the 1983 original.

[12] H. Dong and M. Tao. On the linear convergence to weak/standard d-stationary points of DCA-based algorithms for structured nonsmooth DC programming. *J. Optim. Theory Appl.*, 189(1):190–220, 2021.

[13] J. S. Dyer. The effects of errors in the estimation of the gradient on the Frank-Wolfe algorithm, with implications for interactive programming. *Operations Res.*, 22(1):160–174, 1974.

[14] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3:95–110, 1956.

[15] R. M. Freund and P. Grigas. New analysis and results for the Frank-Wolfe method. *Math. Program.*, 155(1-2, Ser. A):199–230, 2016.

[16] R. M. Freund, P. Grigas, and R. Mazumder. An extended Frank-Wolfe method with “in-face” directions, and its application to low-rank matrix completion. *SIAM J. Optim.*, 27(1):319–346, 2017.

[17] S. Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Math. Program.*, 173(1-2, Ser. A):431–464, 2019.

[18] J.-y. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for sparse optimization problems. *Math. Program.*, 169(1, Ser. B):141–176, 2018.

[19] G. N. Grapiglia. Quadratic regularization methods with finite-difference gradient approximations. *Computational Optimization and Applications*, pages 1–21, 2022.

[20] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152(1-2, Ser. A):75–112, 2015.

[21] O. Hinder, A. Sidford, and N. Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *Conference on learning theory*, pages 1894–1938, 2020.

[22] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. I*, volume 305 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. Fundamentals.

[23] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13:I–427–I–435, 2013.

[24] K. Khamaru and M. J. Wainwright. Convergence guarantees for a class of non-convex and non-smooth optimization problems. *J. Mach. Learn. Res.*, 20:Paper No. 154, 52, 2019.

[25] I. V. Konnov. Simplified versions of the conditional gradient method. *Optimization*, 67(12):2275–2290, 2018.

[26] S. Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[27] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. *arXiv e-prints*, arXiv:1511.05932, 2015.

[28] G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv e-prints*, page arXiv:1309.5550, 2013.

[29] G. Lan and Y. Zhou. Conditional gradient sliding for convex optimization. *SIAM J. Optim.*, 26(2):1379–1409, 2016.

[30] H. A. Le Thi, V. N. Huynh, and T. Pham Dinh. Convergence analysis of difference-of-convex algorithm with subanalytic data. *J. Optim. Theory Appl.*, 179(1):103–126, 2018.

[31] E. Levitin and B. Polyak. Constrained minimization methods. *USSR Comput. Math. Math. Phys.*, 6(5):1–50, 1966.

[32] R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Rev.*, 55(1):65–98, 2013.

[33] N. M. Nam, R. B. Rector, and D. Giles. Minimizing differences of convex functions with applications to facility location and clustering. *J. Optim. Theory Appl.*, 173(1):255–278, 2017.

[34] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1, Ser. A):177–205, 2006.

[35] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[36] F. Pedregosa, G. Negiar, A. Askari, and M. Jaggi. Linearly convergent frank-wolfe with backtracking line-search. *International Conference on Artificial Intelligence and Statistics*, pages 1–10, 2020.

[37] A. K. Sahu, M. Zaheer, and S. Kar. Towards gradient free and projection free stochastic optimization. *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477, 2019.

[38] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet. A majorization-minimization approach to the sparse generalized eigenvalue problem. *Machine learning*, 85(1-2):3–39, 2011.

[39] P. D. Tao and L. T. H. An. Convex analysis approach to d.c. programming: theory, algorithms and applications. *Acta Math. Vietnam.*, 22(1):289–355, 1997.

[40] P. D. Tao and E. B. Souad. Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In *FERMAT days 85: mathematics for optimization (Toulouse, 1985)*, volume 129 of *North-Holland Math. Stud.*, pages 249–271. North-Holland, Amsterdam, 1986.

[41] W. van Ackooij and W. de Oliveira. Nonsmooth and nonconvex optimization via approximate difference-of-convex decompositions. *J. Optim. Theory Appl.*, 182(1):49–80, 2019.

[42] J.-K. Wang and A. Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. *arXiv preprint arXiv:2302.07851*, 2023.

[43] A. Yurtsever and S. Sra. Cccp is frank-wolfe in disguise. *arXiv preprint arXiv:2206.12014*, 2022.