OJMO

Open Journal of Mathematical Optimization

Welington de Oliveira

Short Paper - A note on the Frank–Wolfe algorithm for a class of nonconvex and nonsmooth optimization problems

Volume 4 (2023), article no. 2 (10 pages)

https://doi.org/10.5802/ojmo.21

Article submitted on November 26, 2021, revised on July 30, 2022,
accepted on December 2, 2022.

CENTRE MERSENNE

Open Journal of Mathematical Optimization is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

Short Paper - A note on the Frank–Wolfe algorithm for a class of nonconvex and nonsmooth optimization problems

Welington de Oliveira
Mines Paris, Université PSL, Centre de Mathématiques Appliquées (CMA), 06904 Sophia Antipolis, France
welington.oliveira@minesparis.psl.eu

Abstract

Frank and Wolfe's celebrated conditional gradient method is a well-known tool for solving smooth optimization problems for which minimizing a linear function over the feasible set is computationally cheap. However, when the objective function is nonsmooth, the method may fail to compute a stationary point. In this work, we show that the Frank–Wolfe algorithm can be employed to compute Clarke-stationary points for nonconvex and nonsmooth optimization problems consisting of minimizing upper-$C^{1,\alpha}$ functions over convex and compact sets. Furthermore, under more restrictive assumptions, we propose a new algorithm variant with stronger stationarity guarantees, namely directional stationarity and even local optimality.

Digital Object Identifier 10.5802/ojmo.21

Keywords Nonsmooth Optimization, Nonconvex Optimization, Frank–Wolfe Algorithm.

1 Introduction

The conditional gradient method [11], also known as Frank–Wolfe (FW) algorithm, is one of the simplest and oldest iterative methods for minimizing a (sufficiently) smooth function over a convex and compact set. Despite its modest convergence rate of $\mathcal{O}(1/\sqrt{k})$ for nonconvex objectives [18], the algorithm is particularly attractive when minimizing linear functions over the feasible set is computationally cheap. That is the case of several modern large-scale optimization problems from machine learning and data science, which have revitalized the interest and research on first-order optimization methods. We refer the reader to [1, 6, 12, 16, 18] and references therein for recent developments on the FW algorithm.

In the convex setting, it is well known that the FW algorithm may fail to converge to a stationary point if the objective function is nonsmooth [15, 25]. Approaches for coping with such a shortcoming consist of (a) approximating the objective function with a smooth one, obtained by standard smoothing techniques such as the Moreau–Yosida regularization [24] and randomized rules employing probability densities [10]; (b) considering (when possible) an epigraphic reformulation by adding a new variable and moving the source of nonsmoothness to the constraint [9]; (c) assuming stronger assumptions than a simple oracle that provides an arbitrary subgradient, so that a descent direction can be computed by solving a complex subproblem per iteration [4, 19, 25]. Furthermore, when the objective is the sum of two convex functions h and c, with h smooth and c having a simple structure (e.g., piece-wise linear), then the FW algorithm in its generalized form given in [14] (and revisited in many recent publications) is convergent at the cost of solving a more involving subproblem per iteration.

Without relying on smoothing techniques, particular choices of subgradients, restrictive oracles, or other reformulation tricks that are only applicable in certain particular cases, we show that the classic FW algorithm computes Clarke-stationary points for the broad class of nonconvex and nonsmooth problems of the form

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

where $X \neq \emptyset$ is a convex and compact subset of an open set $\mathfrak{O} \subset \Re^n$, and $f : \mathfrak{O} \to \Re$ can be expressed, over X, as a minimum of a compactly parametrized family of α-Hölder smooth functions (see Definition 1 below). Such a class of functions is denoted by upper-$C^{1,\alpha}$ as $-f$ is lower-$C^{1,\alpha}$, a family of functions introduced in [7].

Contributions and organization

Despite nonsmoothness of the objective function, we show that under standard rules for defining stepsizes, the classic FW algorithm applied to (1) computes Clarke-stationary points and possesses a convergence rate of $\mathcal{O}(1/k^{\frac{\alpha}{\alpha+1}})$, matching the one known for the smooth (but nonconvex) setting (take $\alpha = 1$ and compare with [18, Table 2]). Furthermore, for applications in which f is the point-wise minimum of finitely many α-Hölder smooth functions $F_i : \mathfrak{O} \to \Re$ $(i = 1, \ldots, q)$ we propose a new variant of the FW algorithm with stronger stationarity guarantees, namely directional stationarity. The latter is the sharpest kind among the various stationary concepts in nonsmooth and nonconvex optimization. Furthermore, we show that the concept of d-stationarity is equivalent to local optimality when all functions F_i are convex.

The remainder of this work is organized as follows. First, in Section 2 we recall some basic definitions and provide implementable formulations for two stationarity conditions. Next, the FW algorithm is revisited in Section 3 as well as its convergence analysis for the setting under consideration. The new algorithm variant able to compute directional stationary points is presented in Section 4.

Notation

Throughout this work, $\mathfrak{O}$ is an open set of $\Re^n$ and $\emptyset \neq X \subset \mathfrak{O}$ is a convex and compact set. Given a point $x \in \mathfrak{O}$, we denote by $\mathcal{V}_x \subset \mathfrak{O}$ an open neighborhood of x, that is, the set $\mathcal{V}_x := \{y \in \mathfrak{O} : \ \|y - x\| < \delta\}$ for some $\delta > 0$, where $\|\cdot\|$ is the Euclidean norm. We denote by $N_X(x)$ the normal cone to the set X at the point x: $N_X(x) = \{p : \langle p, y - x\rangle \leq 0 \ \ \forall\, y \in X\}$ if $x \in X$ and $N_X(x) = \emptyset$ otherwise. The indicator function is $\mathbf{i}_X(x) = 0$ if $x \in X$ and $+\infty$ otherwise. The notation α is reserved for scalars in $[0, 1]$.

2 Main definitions, subdifferentiability and stationarity

In this section we present some key definitions and stationary conditions.

▶ **Definition 1** ($UC^{1,\alpha}$ functions). *Let $\alpha \in [0, 1]$. A function $f : \mathfrak{O} \to \Re$ is called upper-$C^{1,\alpha}$ (or $UC^{1,\alpha}$ for short) on $\mathfrak{O}$ if on some open neighborhood $\mathcal{V}_{\bar{x}}$ of each $\bar{x} \in \mathfrak{O}$ there exist a non-empty compact set U, a constant $\ell > 0$, and a continuous function $F : \mathcal{V}_{\bar{x}} \times U \to \Re$ that is differentiable with respect to the x-variable, such that*

$$f(x) := \min_{u \in U} F(x, u) \quad \text{for all} \quad x \in \mathcal{V}_{\bar{x}}, \tag{2a}$$

where $\nabla_x F(x, u)$ is jointly continuous and satisfies, for $I(x) := \arg\min_{u \in U} F(x, u)$, the Hölder condition

$$\|\nabla_x F(y, u) - \nabla_x F(x, u)\| \leq \ell \|y - x\|^\alpha \quad \text{for all } x, y \in \mathcal{V}_{\bar{x}} \text{ and } u \in I(y) \cup I(x). \tag{2b}$$

If $\alpha = 0$, then f is said to be upper-C^1 (i.e, $UC^1 = UC^{1,0}$): condition (2b) becomes superfluous and can be omitted. If $\alpha = 1$, then f is upper-C^2 in view of [7, Rem. 3.3]: $F(\cdot, u)$ are indeed of class $\mathcal{C}^2$, with F and its x-partial derivatives of first and second order jointly continuous on $(x, u) \in \mathcal{V}_{\bar{x}} \times U$. A function φ is said to be lower-C^1 (LC^1), or lower-$C^{1,\alpha}$ ($LC^{1,\alpha}$), or lower-C^2 (LC^2) if $f = -\varphi$ is UC^1, or $UC^{1,\alpha}$, or UC^2, respectively. The class of LC^1 functions was introduced in [23], LC^2 in [20], and $LC^{1,\alpha}$ in [7]. The authors of the latter reference show that, for $\alpha \in (0, 1)$, $LC^2 = LC^{1,1} \subsetneq LC^{1,\alpha} \subsetneq LC^{1,0} = LC^1$. It follows from an analogous argument that $UC^2 = UC^{1,1} \subsetneq UC^{1,\alpha} \subsetneq UC^{1,0} = UC^1$. We care to mention that $UC^{1,\alpha}$ forms a broad class containing all the functions that can be expressed as $f = h - c$, with $h : \mathfrak{O} \to \Re$ α-Hölder and $c : \mathfrak{O} \to \Re$ convex (possibly nonsmooth) [7, Prop. 3.5]. In particular, all Difference-of-Convex (DC) functions [8] with h being α-Hölder are upper-$C^{1,\alpha}$; as a result, all concave functions over X are upper-$C^{1,\alpha}$.

Under the compactness assumption on X the local representation (2) can indeed be extended to a common representation (the same function F, set U, and Lipschitz constant ℓ) for all points in a bounded open set containing X [22, Eq. 10(12)]. Hence, once we assume X compact, there exists a bounded open set $\mathfrak{O}' \subset \mathfrak{O}$ containing X such that the following common representation does not represent a loss of generality with respect to Definition 1: $F : \mathfrak{O}' \times U \to \Re$ is jointly continuous,

$$f(x) := \min_{u \in U} F(x, u), \text{ and } \|\nabla_x F(y, u) - \nabla_x F(x, u)\| \leq \ell \|y - x\|^\alpha \text{ for all } x, y \in \mathfrak{O}' \text{ and } u \in I(y) \cup I(x). \tag{3}$$

From now on we will only consider such a representation for f.

Let $f : \mathfrak{O} \to \Re$ be upper-$C^{1,\alpha}$. It follows from definition that f is locally Lipschitz (see [22, Thm. 10.31] for the Lipschitz constant). Therefore, the Clarke-directional derivative

$$f^\circ(x;d) := \limsup_{x' \to x,\ \tau \downarrow 0} \frac{f(x'+\tau d) - f(x')}{\tau}$$

exists and is finite for all $x \in \mathfrak{O}$ in every direction $d \in \Re^n$ [5, Prop. 2.1.1(a)]. Such a mathematical concept permits to define the Clarke subdifferential of f at $x \in \mathfrak{O}$, $\partial^{\mathtt{C}} f(x) := \{g :\ \langle g, d\rangle \leq f^\circ(x;d) \ \text{ for all } \ d \in \Re^n\}$, which is a nonempty, convex and compact subset of $\Re^n$ [5, Prop. 2.1.2(a)]. For $\varphi = -f$, it follows from (3) that $\varphi(x) = \max_{u \in U} -F(x,u)$ for all points $x \in \mathfrak{O}' \supset X$. Given this structure, Theorem 7.3 of [21] asserts that $\partial^{\mathtt{C}}\varphi(x) = \mathtt{co}\,\{-\nabla_x F(x,u) :\ u \in I(x)\}$ for all $x \in \mathfrak{O}'$. Furthermore, Proposition 2.3.1 in [5] ensures that $\partial^{\mathtt{C}} f(x) = -\partial^{\mathtt{C}}\varphi(x)$ (because $f = -\varphi$), and thus

$$\partial^{\mathtt{C}} f(x) = \mathtt{co}\,\{\nabla_x F(x,u) :\ u \in I(x)\} \neq \emptyset \quad \text{for all } \ x \in \mathfrak{O}' \supset X. \tag{4}$$

Furthermore, it follows from [5, Prop. 2.1.2(b)] that, for all $x \in \mathfrak{O}'$,

$$f^\circ(x;d) = \max_{g \in \partial^{\mathtt{C}} f(x)} \langle g, d\rangle, \quad \text{ and thus } \quad f^\circ(x;d) = \max_{u \in I(x)} \langle \nabla_x F(x,u), d\rangle \quad \text{for all } \ d \in \Re^n. \tag{5}$$

The class of $LC^{1,\alpha}$ functions is *Clarke regular* (because LC^1 is; [20, Thm. 1]), that is, the Clarke-directional derivative coincides with the ordinary directional derivative from Convex Analysis

$$f'(x;d) := \lim_{\tau \downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau}.$$

Unfortunately, this is not the case for $UC^{1,\alpha}$ functions as we may have the strict inequality $f^\circ(x;d) > f'(x;d)$.

▶ **Proposition 2.** *Let $f : \mathfrak{O} \to \Re$ be upper-$C^{1,\alpha}$. Then $f'(x;d) = \min_{g \in \partial^{\mathtt{c}} f(x)} \langle g, d\rangle$ for all $x \in \mathfrak{O}$ and $\in \Re^n$.*

Proof. Let $\varphi = -f$. As φ is $LC^{1,\alpha}$ and $x \in \mathfrak{O}$, it holds that $\varphi^\circ(x;d) = \varphi'(x;d)$ for all $d \in \Re^n$ [20, Thm. 1]. Therefore, the limit $\lim_{\tau \downarrow 0}(\varphi(x+\tau d) - \varphi(x))/\tau$ exists. Note that $\varphi'(x;d) = \lim_{\tau \downarrow 0} -\frac{-\varphi(x+\tau d) - (-\varphi(x))}{\tau} = -\lim_{\tau \downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau}$, showing that $f'(x;d)$ also exists and equals $-\varphi^\circ(x;d)$. Then, $f'(x;d) = -\varphi^\circ(x;d) = -\max_{-g \in \partial^{\mathtt{c}}\varphi(x)} \langle -g, d\rangle = -\max_{g \in \partial^{\mathtt{c}} f(x)} \langle -g, d\rangle = \min_{g \in \partial^{\mathtt{c}} f(x)} \langle g, d\rangle$. ◀

Assuming the common representation (3), Proposition 2 and (4) yield that, for all $x \in \mathfrak{O}'$,

$$f'(x;d) = \min_{u \in I(x)} \langle \nabla_x F(x,u), d\rangle \quad \text{for all and } d \in \Re^n. \tag{6}$$

Non-regularity of $UC^{1,\alpha}$ functions is now evident: compare formulæ (5) and (6). Given the two derivatives f° and f', we can define two stationary conditions for (1).

- (C-stationarity.) A point $\bar{x} \in X$ is said to be a C(larke)-stationary for problem (1) if $0 \in \partial^{\mathtt{C}} f(\bar{x}) + N_X(\bar{x})$. In other words, there exists $g \in \partial^{\mathtt{C}} f(\bar{x})$ such that $-g \in N_X(\bar{x})$, i.e., $\langle g, x - \bar{x}\rangle \geq 0$ for all $x \in X$. We can thus say that $\bar{x}$ is C-stationary for (1) if

$$0 = \min_{z \in X} \langle g, z - \bar{x}\rangle \quad \text{for at least one vector} \quad g \in \partial^{\mathtt{C}} f(\bar{x}). \tag{7}$$

  By adopting the representation (3) for f, C-stationarity becomes

$$\min_{z \in X} \langle \nabla_x F(\bar{x}, \bar{u}), z - \bar{x}\rangle = 0 \text{ for at least one index } \bar{u} \in I(\bar{x}). \tag{8}$$

- (d-stationarity.) A point $\bar{x} \in X$ is said to be a d(irectional)-stationary for problem (1) if $f'(\bar{x}; x - \bar{x}) \geq 0$ for all $x \in X$. It follows from Proposition 2 that d-stationarity means

$$0 = \min_{z \in X} \langle g, z - \bar{x}\rangle \quad \text{for all} \quad g \in \partial^{\mathtt{C}} f(\bar{x}). \tag{9}$$

  By adopting the representation (3) for f, d-stationarity becomes

$$0 = \min_{z \in X} \langle \nabla_x F(\bar{x}, \bar{u}), z - \bar{x}\rangle \quad \text{for all} \quad \bar{u} \in I(\bar{x}). \tag{10}$$

Equations (7) and (9) show that d-stationarity is a much stronger condition than C-stationarity. Indeed, d-stationarity is the sharpest kind among the various stationary concepts in nonsmooth and nonconvex optimization [17]. Both conditions coincide when f is smooth at the point under consideration: in this case, $\partial^{\mathsf{C}} f(\bar{x}) = \{\nabla f(\bar{x})\}$.

▶ **Example 3.** Let $f : \Re \to \Re$ be defined as $f(x) = \min\{(x+2)^2, (x-2)^2\}$ and $X = [0, 4]$. Note that f is UC^2 (thus $UC^{1,\alpha}$): $F_1(x) := F(x, 1) = (x+2)^2$, $F_2(x) := F(x, 2) = (x-2)^2$ are C^2 and $U = \{1, 2\}$. Furthermore, f is nondifferentiable at $\bar{x} = 0$: equation (4) gives $\partial^{\mathsf{C}} f(0) = [-4, 4]$, (5) yields $f^\circ(0; \pm 1) = 4$, however, $f'(0; \pm 1) = -4$. Hence f is not regular at $\bar{x} = 0$, a point that is C-stationary (in fact a global maximizer) but not d-stationary. Indeed, for all $g \in [0, 4] \subset \partial^{\mathsf{C}} f(0)$ we get $0 = \min_{x\in[0,4]} g \cdot (x - 0)$.

The following result shows that if a d-stationary point lies in the interior of X, then f is smooth at this point.

▶ **Proposition 4.** *Let $f : \mathfrak{O} \to \Re$ be an upper-$C^{1,\alpha}$ function, and $\bar{x} \in \mathtt{int}(X)$ be a d-stationary point for (1). Then $\partial^{\mathsf{C}} f(\bar{x}) = \{0\}$.*

Proof. Let $g \in \partial^{\mathsf{C}} f(\bar{x})$ be an arbitrary subgradient. Since $\bar{x} \in \mathtt{int}(X)$, there exists $\epsilon > 0$ such that $x = \bar{x} - \epsilon \frac{g}{\|g\|} \in X$. The assumption of d-stationarity implies that $0 \le \langle g, x - \bar{x}\rangle = \langle g, -\epsilon \frac{g}{\|g\|}\rangle = -\epsilon\|g\|$, showing that $g = 0$. ◀

A word of caution may be necessary: the above result does not imply that the index set $I(\bar{x})$ is a singleton. It implies that $\nabla_x F(\bar{x}, \bar{u}) = 0$ for all $\bar{u} \in I(\bar{x})$. (As an example, take $f(x) = \min\{x^2, 2x^2\}$ and $\bar{x} = 0$.)

3 Revisiting the Frank–Wolfe algorithm for $UC^{1,\alpha}$ functions

The alternative representation of C-stationarity condition given by (7) motivates us to apply the FW algorithm of [11] to problem (1). In Algorithm 1 we assume the existence of an oracle that, for any given point $x \in X$, provides us with the value $f(x)$ and an arbitrary Clarke-subgradient $g \in \partial^{\mathsf{C}} f(x)$.

Algorithm 1 The Frank–Wolfe Algorithm

1: Let $x^0 \in X$ and $\mathtt{Tol} \ge 0$ be given
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Call the oracle to obtain an arbitrary subgradient $g^k \in \partial^{\mathsf{C}} f(x^k)$ and compute $z^k \in \arg\min_{z\in X} \langle g^k, z\rangle$
4: Set $d^k = z^k - x^k$ and $\theta_k = -\langle g^k, d^k\rangle$
5: Stop if $\theta_k \le \mathtt{Tol}$: x^k is a C-stationary point within tolerance $\mathtt{Tol}$
6: Choose $\tau_k \in (0, 1]$ and set $x^{k+1} = x^k + \tau_k d^k$
7: **end for**

Algorithm 1 is interesting in situations where minimizing a linear function over X is computationally cheap. We refer the interested reader to [12] and references therein for several optimization problems arising from machine learning and data science that are suitable for the FW algorithm. Differently from [12] that deals with smooth problems, the recent work [19] also gives a list of nonsmooth but convex optimization problems that can be solved by a variant of the FW algorithm tailored to a particular structure of nonsmoothness. The main idea in [19], that dates back to [25], is to work with well-chosen approximate subgradients of f. More precisely, instead of computing an arbitrary subgradient, the method requires at every iteration the construction of a set $T_k \subset \Re^n$ containing all the subgradients of f in a neighborhood of x^k and defines z^k by solving the more involving subproblem $\min_{z\in X} \max_{s\in T_k} \langle s, z\rangle$, which is implementable only in some particular (convex) cases [4, 19, 25].

Under the assumption that f is smooth, then Algorithm 1 is the classic conditional gradient method of [11]. The sole contribution of this section is the proof that the algorithm (asymptotically) computes a C-stationary point of (1) provided f is $UC^{1,\alpha}$ with $\alpha \in (0, 1]$. To this end, we assume that $\mathtt{Tol} = 0$ and the algorithm does not stop. (If Algorithm 1 stops at iteration k, then $\theta_k = 0$ because $\theta_k \ge 0$ for all k: in this case, $0 = \min_{z\in X} \langle g^k, z - x^k\rangle$ and thus x^k is C-stationary; cf. (7).) We start our analysis with the following key result. Once the inequality in Lemma 5 is established, convergence analysis follows the same techniques found in the (smooth) FW algorithm's literature.

▶ **Lemma 5.** *Consider problem (1) with $f : \mathfrak{O} \to \Re$ satisfying (3), $\alpha \in (0, 1]$, and $X \subset \mathfrak{O}$ a convex and compact set. Then, for the (possibly unknown) constant ℓ given in (3)*

$$f(x^{k+1}) \le f(x^k) - \tau_k\theta_k + \frac{\ell}{\alpha+1}\tau_k^{\alpha+1}\|d^k\|^{\alpha+1}, \quad \forall\, k = 0, 1, 2\ldots$$

Proof. Let $I(x^k) = \{u \in U : \ F(x^k, u) = f(x^k)\}$. It follows from the Hölder condition in (3) (see also [15, Eqs. (2.4) and (2.5)]) that, for all $u \in I(x^k)$, $F(x^{k+1}, u) \leq F(x^k, u) + \langle \nabla F(x^k, u), x^{k+1} - x^k \rangle + \frac{\ell}{\alpha+1}\|x^{k+1} - x^k\|^{\alpha+1}$. The Carathéodory Theorem [22, Thm. 2.29] ensures that every $g \in \partial^{\mathsf{C}} f(x^k) = \mathtt{co}\left\{\nabla_x F(x^k, u) \ : \ u \in I(x^k)\right\}$ can be written as a convex combination of no more than $n+1$ vectors $\nabla_x F(x^k, u)$, $u \in I(x^k)$. Therefore, by replicating $u_i \in I(x^k)$ and assigning $\lambda_i^k = 0$ if necessary, the subgradient g^k at iteration k of Algorithm 1 can be expressed as $g^k = \sum_{i=1}^{n+1} \lambda_i^k \nabla_x F(x^k, u_i)$,with $\lambda^k \in \Re_+^{n+1}$ s.t. $\sum_{i=1}^{n+1} \lambda_i^k = 1$, and $u_i \in I(x^k)$. We get from the above inequality that, for $u_i \in I(x^k)$,

$$\sum_{i=1}^{n+1} \lambda_i^k F(x^{k+1}, u_i) \leq \sum_{i=1}^{n+1} \lambda_i^k F(x^k, u_i) + \langle g^k, x^{k+1} - x^k \rangle + \frac{\ell}{\alpha+1}\|x^{k+1} - x^k\|^{\alpha+1}.$$

Recall that $\sum_{i=1}^{n+1} \lambda_i^k F(x^k, u_i) = f(x^k)$ because $u_i \in I(x^k)$. Furthermore,

$$\sum_{i=1}^{n+1} \lambda_i^k F(x^{k+1}, u_i) \geq \sum_{i=1}^{n+1} \lambda_i^k \min_{u \in U} F(x^{k+1}, u) = f(x^{k+1}).$$

We have thus shown that $f(x^{k+1}) \leq f(x^k) + \langle g^k, x^{k+1} - x^k \rangle + \frac{\ell}{\alpha+1}\|x^{k+1} - x^k\|^{\alpha+1}$. The result follows from definitions $x^{k+1} = x^k + \tau_k d^k$ and $\theta_k = -\langle g^k, d^k \rangle$. ◀

The next two theorems follow from simple adjustments (to the configuration under consideration) of known results found in the FW algorithm's literature for smooth problems, see e.g. [2, Chap. 13] and [13, §B.2.1]. For this reason, we move their proofs to the Appendix.

▶ **Theorem 6.** *Under the setting of Lemma 5, let* $\underline{\theta}_k := \min_{j \leq k} \theta_j$ *and* $f^* := \min_{x \in X} f(x)$. *Then,*

$$\underline{\theta}_k \leq \frac{f(x^0) - f^* + \frac{\ell}{\alpha+1}\mathcal{D}\mathrm{iam}(X)^{\alpha+1} \sum_{j=0}^{k} \tau_j^{\alpha+1}}{\sum_{j=0}^{k} \tau_j}. \tag{11}$$

If $\{\tau_k\}$ *satisfies* $\sum_{k=0}^{\infty} \tau_k = \infty$ *and* $\lim_{k\to\infty} \tau_k = 0$, *then* $\lim_{k\to\infty} \underline{\theta}_k = 0$. *Furthermore, let* $j(k) \in \{1, \ldots, k\}$ *be such that* $\underline{\theta}_k = \theta_{j(k)}$. *Then any cluster point of the sequence* $\{x^{j(k)}\}$ *is a C-stationary point for* (1).

The sequence $\{f(x^k)\}$ can be made monotone upon more strict rules to define stepsizes. The following result is an adaptation of [2, Thm. 13.9] (that considers $\alpha = 1$ and f to be ℓ-smooth).

▶ **Theorem 7.** *Consider Algorithm 1 with* $\mathtt{Tol} = 0$ *applied to problem* (1), *with* $f : \mathfrak{O} \to \Re$ *satisfying* (3), $\alpha \in (0, 1]$, *and* $X \subset \mathfrak{O}$ *a convex and compact set. Again, let* $\underline{\theta}_k := \min_{j \leq k} \theta_j$ *and* $f^* := \min_{x \in X} f(x)$. *Then, under the stepsize rule (i)* $\tau_k = \min\left\{\left(\frac{\theta_k}{\ell \|d^k\|^{\alpha+1}}\right)^{\frac{1}{\alpha}}, 1\right\}$ *or (ii)* $\tau_k \in \arg\min_{\tau \in [0,1]} f(x^k + \tau d^k)$, *we have that*

$$0 \leq \underline{\theta}_k \leq \begin{cases} \frac{\frac{\alpha+1}{\alpha}[f(x^0) - f^*]}{k+1} & \text{if } k < \left(\frac{\frac{\alpha+1}{\alpha}[f(x^0) - f^*]}{\ell \mathcal{D}\mathrm{iam}(X)^{\alpha+1}}\right) - 1 \\ \left(\frac{\frac{\alpha+1}{\alpha}[f(x^0) - f^*]\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}{k+1}\right)^{\frac{\alpha}{\alpha+1}} & \text{otherwise.} \end{cases}$$

In particular, $\lim_{k\to\infty} \underline{\theta}_k = 0$ *at rate of* $\mathcal{O}\left(\frac{1}{k^{\frac{\alpha}{\alpha+1}}}\right)$.

This result is a conceptual one because the Lipschitz constant ℓ in (i) is in general unknown and rule (ii) amounts to globally solving a uni-dimensional function over the interval $[0, 1]$. Less stringent schemes that work well in practice employ inexact line-searches (e.g. [18] and the Armijo rule).

Let us consider again Example 3, and apply Algorithm 1. If we start with $x^0 = 0$ and the oracle returns $g^0 = \nabla F_1(x^0) = 4$, then we get $\theta_0 = 0$: the algorithm stops at iteration $k = 0$ with the C-stationary point x^0 that is a global maximizer. This fact motivates the following section.

4 Directional stationarity via a modified FW algorithm

Ideally, for nonsmooth minimization problems, one wants to design an algorithm that will compute a stationary point that has the best chance to be a local minimum [17]. However, without further structure, having guarantees of computing a d-stationary point is out of reach. One special structure arises when X has finitely many known vertices, a setting exploited in [3] for more structured functions and briefly adapted to our framework in the

Appendix. We now propose a new variant of the FW algorithm for computing d-stationary points upon the following additional assumption on the objective function: f is given by

$$f(x) = \min_{i=1,\dots,q} F_i(x), \quad \text{with every } \; F_i : \mathfrak{O} \to \Re \quad \text{known and } \alpha\text{-Hölder smooth}, \tag{12}$$

i.e, U in (3) is the finite index set $\{1,\dots,q\}$. Accordingly, it follows from (9) that $\bar{x} \in X$ is a d-stationarity point of problem (1) if

$$0 = \min_{x \in X} \langle \nabla F_i(\bar{x}), x - \bar{x} \rangle \quad \text{for all } \; i \in I(\bar{x}) := \{j \leq q : \; F_j(\bar{x}) = f(\bar{x})\}. \tag{13}$$

Based on this fact, Algorithm 2 seeks, at every iteration, for a descent direction by checking the gradient of all functions F_i that are ϵ-active. To be more precise, we define the following index set, with $\epsilon > 0$ a small tolerance $I_\epsilon(x) = \{i \leq q : \; f(x) \geq F_i(x) - \epsilon\}$. We say that F_i is ϵ-active at x if $i \in I_\epsilon(x)$. If $q = 1$ in (12), then

Algorithm 2 d-stationary Frank–Wolfe Algorithm

Let $x^0 \in X$, $\texttt{Tol} \geq 0$, and $\epsilon > 0$ be given
for $k = 0, 1, 2, \dots$ **do**
  for $i \in I_\epsilon(x^k)$ **do**
    Let $z^{i,k} \in X$ be a solution of $\min_{x \in X} \langle \nabla F_i(x^k), x - x^k \rangle$
    Set $d^{i,k} = z^{i,k} - x^k$. Choose $\tau_{i,k} \in (0, 1]$ if $\langle \nabla F_i(x^k), d^{i,k} \rangle < 0$; otherwise $d^{i,k} = 0$ and $\tau_{i,k} = 0$
    Define $x^{i,k} = x^k + \tau_{i,k} d^{i,k}$ and $\theta_i(x^k) = -\langle \nabla F_i(x^k), d^{i,k} \rangle$
  end for
  Stop if $\max_{i \in I(x^k)} \theta_i(x^k) \leq \texttt{Tol}$: x^k is a d-stationary point within tolerance Tol
  Define $i^* \in \arg\min_{i \in I_\epsilon(x^k)} f(x^{i,k})$ and set $x^{k+1} = x^{i^*,k}$
end for

Algorithm 2 boils down to the classic FW algorithm. Again aligned with the main motivation from [4, 19, 25], when $q > 1$ Algorithm 2 searches for a subgradient in $\partial^{\mathsf{C}} f$ yielding maximum descent. In order to provide an asymptotic analysis, we must allow the possibility of employing "approximate" subgradients yielded by $\nabla F_i(x^k)$ with i in $I_\epsilon(x^k) \backslash I(x^k)$. This is in the same vein as the proximal method of [17] for DC programming. Note that $\theta_i(x)$ is nonnegative for all $i \in I_\epsilon(x)$ and all $x \in X$. Naturally, if at iteration k we have $\theta_i(x^k) = 0$ for all active index $i \in I(x^k)$, then $\bar{x} = x^k$ is, from (13), d-stationary for problem (1) and the algorithm should halt. This explains why the stopping test above employs the set $I(x^k)$ instead of $I_\epsilon(x^k)$. Indeed, the alternative stopping test $\max_{i \in I_\epsilon(x^k)} \theta_i(x^k) \leq \texttt{Tol}$ may never be triggered even when $\texttt{Tol} > 0$: we may find a direction $d^{j,k}$ that is of descent for some F_j with $j \in I_\epsilon(x^k) \backslash I(x^k)$ but not for f.

In what follows we analyze the convergence properties of Algorithm 2. To this end, we need to assert on the continuity of the function $\theta_i(x)$. Let $\omega_i : \mathfrak{O} \times \mathfrak{O} \to \Re$ be given by $\omega_i(x, y) := \langle \nabla F_i(y), x - y \rangle$. Since ∇F_i is continuous by assumption, $\omega_i(x, y)$ is continuous on both arguments. Then, Theorem 1.17(c) from [22] ensures that $\theta_i(y) = -\min_{x \in X} \; \omega_i(x, y)$ is a continuous function.

▶ **Theorem 8.** *Let $f : \mathfrak{O} \to \Re$ be given by* (12) *and $X \subset \mathfrak{O}$ a convex and compact set. Consider Algorithm 2 with $\epsilon > 0$ applied to problem* (1) *and suppose that the sequence of stepsizes is defined by one of the following rules*[1] *(i) $\tau_{i,k} = \min\left\{ \left(\frac{\theta_i(x^k)}{\ell \|d^{i,k}\|^{\alpha+1}} \right)^{\frac{1}{\alpha}}, 1 \right\}$ or (ii) $\tau_{i,k} \in \arg\min_{\tau \in [0,1]} f(x^k + \tau d^{i,k})$. Then every cluster point of the sequence $\{x^k\}$ generated by the algorithm is d-stationary for* (1).

Proof. Let us define $G_i(x^k) = \frac{\alpha}{\alpha+1} \theta_i(x^k) \min\left\{ \frac{\theta_i(x^k)^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}} \mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}, 1 \right\} \geq 0$. Clearly, $G_i(x^k) = 0$ if, and only if, $\theta_i(x^k) = 0$. If $\theta_i(x^k) > 0$ we can proceed as in Lemma 12 (Appendix) with f replaced by F_i to conclude that $F_i(x^{i,k}) \leq F_i(x^k) - G_i(x^k)$ under both rules (i) and (ii). If $\theta_i(x^k) = 0$, then $G_i(x^k) = 0$. In both cases, $F_i(x^k) - G_i(x^k) \geq F_i(x^{i,k}) \geq f(x^{i,k}) \geq f(x^{k+1})$ for all $i \in I_\epsilon(x^k)$, showing that the sequence $\{f(x^k)\}$ is nonincreasing (because $F_i(x^k) = f(x^k)$ for $i \in I(x^k)$). Let $\bar{x} \in X$ be an arbitrary cluster point of $\{x^k\}$ and $\{x^{k_l}\}$ be a subsequence such that $\lim_{l \to \infty} x^{k_l} = \bar{x}$. It follows from continuity of F_i, $i = 1, \dots, q$, that $I(\bar{x}) \subset I_\epsilon(x^{k_l})$ for all l large enough. Then, by working with such indexes we obtain $F_i(x^{k_l}) - G_i(x^{k_l}) \geq f(x^{k_l+1}) \geq \cdots \geq f(x^{k_{l+1}})$ for all $i \in I(\bar{x})$. By taking the limit with l going to infinite we get from continuity that $F_i(\bar{x}) - f(\bar{x}) \geq \lim_{l \to \infty} G_i(x^{k_l}) = G_i(\bar{x}) \geq 0$ for all $i \in I(\bar{x})$. Recall that $F_i(\bar{x}) = f(\bar{x})$ for $i \in I(\bar{x})$ and thus $G_i(\bar{x}) = 0$.

[1] $\ell > 0$ is the maximum among the Lipschitz constants of functions F_i, $i = 1, \dots, q$.

Furthermore, it follows from the definition of G_i that $0 = \theta_i(\bar{x}) = -\min_{x\in X}\langle \nabla F_i(\bar{x}), x - \bar{x}\rangle$ for all $i \in I(\bar{x})$, i.e., $\bar{x}$ is a d-stationary point. ◀

We highlight that the convergence rate of Theorem 7 applies.

▶ **Example 9.** Let $F_1(x) := F(x,1) = x^2/2 + x$, $F_2(x) := F(x,2) = x^2/2$, and $f(x) = \min\{F_1(x), F_2(x)\}$. As for the feasible set, let us take $X = [-4, 2]$. Suppose we employ Algorithm 1 with $x^0 = -3$ and $\tau_0 = 3/5$. In this case, $f(-3) = F_1(-3) = 3/2$ and $\nabla F_1(-3) = -2$. It is easy to see that the algorithm defines $d^0 = 2 - (-3) = 5$ and, thus, $x^1 = x^0 + \tau_0 d^0 = -3 + (3/5)5 = 0$. At this point, $f(0) = F_1(0) = F_2(0) = 0$ and f is non-differentiable: $\partial^{\mathtt{C}} f(0) = [0, 1]$. If the oracle returns $g^1 = \nabla F_2(0) = 0 \in \partial^{\mathtt{C}} f(x^1)$, then $\theta_1 = 0$ and the algorithm stops at the C-stationary point $x^1 = 0$ after one iteration. It is clear that x^1 is not d-stationary (f is non-differentiable at $x^1 \in \mathtt{int}(X)$, cf. Prop. 4). Suppose now we employ Algorithm 2 with the same starting point and $\epsilon = 0.1$. In this case, the first iteration coincides with that of Algorithm 1 (because $I_\epsilon(x^0) = I(x^0) = \{1\}$). Thus $x^1 = 0$ and $I_\epsilon(x^1) = I(x^1) = \{1, 2\}$. Algorithm 2 considers the two subgradients (of f) $\nabla F_1(0) = 1$ and $\nabla F_2(0) = 0$. With the first choice, the algorithm computes a descent direction and escapes from the C-stationary point $x^1 = 0$.

The next result shows how sharp the d-stationary concept is for the problems of interest.

▶ **Theorem 10.** *Let $\bar{x} \in X$ be d-stationary for (1) with f given by (12). In addition, assume that there exists $\delta > 0$ such that F_i in (12) is convex over $B(\bar{x}, \delta)$ for all $i \in I(\bar{x})$. Then $\bar{x}$ is a local solution of problem (1).*

Proof. Since $\bar{x}$ is d-stationary, if all functions F_i are active at $\bar{x}$, that is, $\{1, \ldots, q\} = I(\bar{x})$, then $\langle \nabla F_i(\bar{x}), x - \bar{x}\rangle \geq 0$ for all $x \in X$ and all $i \in \{1, \ldots, q\}$. In particular, $\langle \nabla F_i(\bar{x}), x - \bar{x}\rangle \geq 0$ for all $x \in X \cap B(\bar{x}, \delta)$ and the local convexity of F_i yields that $\bar{x}$ minimizes F_i over $X \cap B(\bar{x}, \delta)$. Hence, $F_i(\bar{x}) \leq F_i(x)$ for all $x \in X \cap B(\bar{x}, \delta)$ and all $i \in \{1, \ldots, q\}$, resulting in $f(\bar{x}) \leq f(x)$ for all $x \in X \cap B(\bar{x}, \delta)$, i.e., $\bar{x}$ is a local solution of (1).

Suppose now that $I(\bar{x}) \subsetneq \{1, \ldots, q\}$, and let $j \in \{1, \ldots, q\}\backslash I(\bar{x})$, i.e., $F_j(\bar{x}) > f(\bar{x})$. Set $\epsilon_j = (F_j(\bar{x}) - f(\bar{x}))/2 > 0$. Continuity of F_j ensures that there exists $\delta_j > 0$ such that $x \in B(\bar{x}, \delta_j)$ implies $F_j(x) \geq F_j(\bar{x}) - \epsilon_j > f(\bar{x})$. As there are only finitely many functions F_j, we conclude that $\bar{\epsilon} = \min_{j\in\{1,\ldots,q\}\backslash I(\bar{x})} \epsilon_j$ is a strictly positive constant. Again, continuity ensures the existence of $\widetilde{\delta} > 0$ such $x \in B(\bar{x}, \widetilde{\delta})$ implies $F_j(x) \geq F_j(\bar{x}) - \bar{\epsilon} > f(\bar{x})$ for all $j \in \{1, \ldots, q\}\backslash I(\bar{x})$. Moreover, convexity of F_i over $B(\bar{x}, \delta)$ for $i \in I(\bar{x})$, and d-stationarity of $\bar{x}$ yields that $F_i(x) \geq F_i(\bar{x}) = f(\bar{x})$ for all $x \in X \cap B(\bar{x}, \delta)$. Let $\bar{\delta} = \min\{\delta, \widetilde{\delta}\} > 0$. By combining the last two inequalities we conclude that, for all $x \in X \cap B(\bar{x}, \bar{\delta})$, $F_i(x) \geq f(\bar{x})$ for all $i \in \{1, \ldots, q\}$, which gives $f(x) \geq f(\bar{x})$ for all $x \in X \cap B(\bar{x}, \bar{\delta})$, i.e., $\bar{x}$ is a local solution for (1). ◀

▶ **Corollary 11.** *If all F_i in (12) are convex, then every d-stationary point of problem (1) is a local solution.*

This corollary, apparently innocuous, conveys an original result in the area of DC programming [8]. Indeed, if all functions F_i in (12) are convex, then f can be decomposed as a DC function $f(x) = f_1(x) - f_2(x)$, with $f_1(x) = \sum_{i=1}^m F_i(x)$ and $f_2(x) = \max_{j=1,\ldots,m} \sum_{\ell\neq j} F_\ell(x)$. In this setting, the concept of d-stationarity of a point $\bar{x} \in X$ to problem $\min_{x\in X} f_1(x) - f_2(x)$ is equivalent to the inclusion $\partial f_2(\bar{x}) \subset \partial[f_1(\bar{x}) + \mathbf{i}_X(\bar{x})]$. Hence, all local minimizers must satisfy this condition. However, the reverse implication is only known to hold if f_2 is (locally) polyhedral. Corollary 11 gives another framework where d-stationarity implies local optimality without needing f_2 to be locally polyhedral.

Acknowledgments

The author thanks the reviewers for their constructive comments, which helped to improve this work significantly. The author also acknowledges financial support from the Gaspard–Monge Program for Optimization and Operations Research (PGMO) project "Scalable Optimization for Learning and Energy Management."

A Appendix

A.1 Proof of Theorem 6.

It follows from Lemma 5 (and the fact that $\|d^k\|^{\alpha+1} = \|z^k - x^k\|^{\alpha+1} \leq \mathcal{D}\mathrm{iam}(X)^{\alpha+1}$) that $\sum_{j=0}^k \tau_j\theta_j \leq \sum_{j=0}^k [f(x^j) - f(x^{j+1}) + \frac{\ell}{\alpha+1}\tau_j^{\alpha+1}\mathcal{D}\mathrm{iam}(X)^{\alpha+1}] = f(x^0) - f(x^{k+1}) + \frac{\ell}{\alpha+1}\mathcal{D}\mathrm{iam}(X)^{\alpha+1}\sum_{j=0}^k \tau_j^{\alpha+1}$. Since $f^* \leq f(x^k)$ for all k, (11) holds. Our assumptions on $\{\tau_k\}$ yields $\big(\sum_{j=0}^k \tau_j^{\alpha+1}\big)/\big(\sum_{j=0}^k \tau_j\big) \to 0$ as $k \to \infty$ [13, §B.2.1] and thus $0 \leq \lim_{k\to\infty} \underline{\theta}_k \leq \lim_{k\to\infty} \big(f(x^0) - f^* + \frac{\ell}{\alpha+1}\mathcal{D}\mathrm{iam}(X)^{\alpha+1}\sum_{j=0}^k \tau_j^{\alpha+1}\big)/\big(\sum_{j=0}^k \tau_j\big) = 0$. Furthermore,

let $j(k) \in \{1,\dots,k\}$ be s.t. $\underline{\theta}_k = \theta_{j(k)}$. Then, $\langle g^{j(k)}, x - x^{j(k)}\rangle \geq \langle g^{j(k)}, d^{j(k)}\rangle = -\theta_{j(k)} = -\underline{\theta}_k$ for all $x \in X$. Recall that $\partial^{\mathcal{C}} f$ is closed and locally bounded over $\mathfrak{D}$. It thus follows from boundedness of $\{x^{j(k)}\} \subset X$ that $g^{j(k)}$ is bounded as well. Let $\bar{x} \in X$ be an arbitrary cluster point of $\{x^{j(k)}\}$; by extracting subsequences if necessary, we get two convergent subsequences $\{x^{j(k')}\}$ and $\{g^{j(k')}\}$: $x^{j(k')} \to \bar{x} \in X$ and $g^{j(k')} \to \bar{g} \in \partial^{\mathcal{C}} f(\bar{x})$ [5, Prop. 2.1.5(c)]. It is clear from the above development that $\langle g^{j(k')}, x - x^{j(k')}\rangle \geq -\underline{\theta}_{k'}$ for all $x \in X$. By passing to the limit as k' goes to infinity, and recalling that $\underline{\theta}_{k'} \to 0$, we conclude that $\langle \bar{g}, x - \bar{x}\rangle \geq 0$ for all $x \in X$. Hence, it follows from (7) that $\bar{x} \in X$ is a C-stationary point for problem (1).

A.2 Proof of Theorem 7.

We first prove the following auxiliary result, which is an adaption of [2, Lem. 13.8].

▶ **Lemma 12.** *Under the assumptions of Lemma 5, suppose that the sequence of stepsizes is defined by one of the following rules (i)* $\tau_k = \min\left\{\left(\frac{\theta_k}{\ell\|d^k\|^{\alpha+1}}\right)^{\frac{1}{\alpha}}, 1\right\}$ *or (ii)* $\tau_k \in \arg\min_{\tau\in[0,1]} f(x^k + \tau d^k)$*. Then the sequence of function values satisfies*

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{\alpha+1}\theta_k \min\left\{\frac{\theta_k^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}, 1\right\} \quad \forall\ k = 0,1,2\dots \tag{14}$$

Proof. Let us define the following uni-dimensional function $w(\tau) := f(x^k) - \tau\theta_k + \frac{\ell}{\alpha+1}\tau^{\alpha+1}\|d^k\|^{\alpha+1}$, which is well defined and strictly convex over $[0,\infty)$. By solving the problem $\min_{\tau\in[0,\infty)} w(\tau)$ we get as solution the point $\bar{\tau}_k$ satisfying $0 = -\theta_k + \ell\bar{\tau}_k^{\alpha}\|d^k\|^{\alpha+1}$, i.e., $\bar{\tau}_k = (\frac{\theta_k}{\ell\|d^k\|^{\alpha+1}})^{\frac{1}{\alpha}} > 0$. By evaluating w at $\bar{\tau}_k$ we get

$$\begin{aligned} w(\bar{\tau}_k) &= f(x^k) - \frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}} + \frac{\ell}{\alpha+1}\frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{\alpha+1}{\alpha}}\|d^k\|^{\frac{(\alpha+1)^2}{\alpha}}}\|d^k\|^{\alpha+1} \\ &= f(x^k) - \frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}} + \frac{1}{\alpha+1}\frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}} = f(x^k) - \frac{\alpha}{\alpha+1}\frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}}. \end{aligned}$$

We now take $\tau_k = \min\{\bar{\tau}_k, 1\}$ for all k, which corresponds to rule (i), and analyze the following two cases:

- $\tau_k < 1$. In this case, $\tau_k = \bar{\tau}_k$ and thus $f(x^{k+1}) \leq w(\tau_k) = f(x^k) - \frac{\alpha}{\alpha+1}\frac{\theta_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}}$ from Lemma 5.
- $\tau_k = 1$. Then $f(x^{k+1}) \leq w(\tau_k) = w(1) = f(x^k) - \theta_k + \frac{\ell}{\alpha+1}\|d^k\|^{\alpha+1}$. Observe that $\tau_k = 1$ implies $\theta_k^{\frac{1}{\alpha}} \geq (\ell\|d^k\|^{\alpha+1})^{\frac{1}{\alpha}}$, which in turn gives $\theta_k \geq \ell\|d^k\|^{\alpha+1}$ because all these variables and parameters are non-negative. Therefore, $f(x^{k+1}) \leq f(x^k) - \theta_k + \frac{\ell}{\alpha+1}\|d^k\|^{\alpha+1} \leq f(x^k) - \frac{\alpha}{\alpha+1}\theta_k$.

In both cases we have

$$f(x^{k+1}) \leq w(\tau_k) = f(x^k) - \frac{\alpha}{\alpha+1}\theta_k \min\left\{\frac{\theta_k^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\|d^k\|^{\frac{\alpha+1}{\alpha}}}, 1\right\} \leq f(x^k) - \frac{\alpha}{\alpha+1}\theta_k \min\left\{\frac{\theta_k^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}, 1\right\},$$

as stated. The analysis for item (ii) is straightforward: as (14) holds for the non-optimal rule of item (i), then it must hold for the rule of item (ii) because $f(x^{k+1}) = \min_{\tau\in[0,1]} f(x^k + td^k) \leq f(x^k + \min\{\bar{\tau}_k, 1\}d^k)$. ◀

We now proceed to the proof of Theorem 7. Regardless the stepsize rule (i) or (ii), Lemma 12 ensures that $f^* - f(x^0) \leq f(x^{k+1}) - f(x^0) = \sum_{j=0}^{k}[f(x^{j+1}) - f(x^j)] \leq -\sum_{j=0}^{k}\frac{\alpha}{\alpha+1}\theta_k \min\left\{\frac{\theta_k^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}, 1\right\}$, showing that

$$(k+1)\underline{\theta}_k \min\left\{\frac{\underline{\theta}_k^{\frac{1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}, 1\right\} \leq \frac{\alpha+1}{\alpha}[f(x^0) - f^*]. \tag{15}$$

Suppose that $\underline{\theta}_k^{\frac{1}{\alpha}} > \ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}$. Then (15) yields $\underline{\theta}_k \leq \frac{\frac{\alpha+1}{\alpha}[f(x^0)-f^*]}{k+1}$. Combining these two last inequalities we conclude that the former cannot hold for $k \geq \left(\frac{\frac{\alpha+1}{\alpha}[f(x^0)-f^*]}{\ell\mathcal{D}\mathrm{iam}(X)^{\alpha+1}}\right) - 1$. Therefore, after finitely many steps we have that $\underline{\theta}_k^{\frac{1}{\alpha}} \leq \ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}$ for all k large enough. In this case, (15) gives $\frac{\underline{\theta}_k^{\frac{\alpha+1}{\alpha}}}{\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}} \leq \frac{\frac{\alpha+1}{\alpha}[f(x^0)-f^*]}{k+1}$, i.e., $\underline{\theta}_k \leq \left(\frac{\frac{\alpha+1}{\alpha}[f(x^0)-f^*]\ell^{\frac{1}{\alpha}}\mathcal{D}\mathrm{iam}(X)^{\frac{\alpha+1}{\alpha}}}{k+1}\right)^{\frac{\alpha}{\alpha+1}}$. This inequality proves that $\lim_{k\to\infty}\underline{\theta}_k = 0$ at convergence rate of $\mathcal{O}\left(\frac{1}{k^{\frac{\alpha}{\alpha+1}}}\right)$. ◀

A.3 Computing d-stationary points when the feasible set has known vertices

▶ **Proposition 13.** *Let $f : \mathfrak{O} \to \Re$ be an upper-$C^{1,\alpha}$ function and suppose that X has finitely many vertices $v^1 \ldots, v^m$. Then $\bar{x}$ is a d-stationary point of problem* (1) *if and only if $f'(\bar{x}; v^\iota - \bar{x}) \geq 0$ for all $\iota = 1, \ldots, m$.*

Proof. The first implication follows directly from the definition of d-stationarity: $f'(\bar{x}; x - \bar{x}) \geq 0$ for all $x \in X$. To prove the converse implication, note that for every $x \in X = \mathtt{co}\left\{v^1 \ldots, v^m\right\}$, there exists a vector $\lambda_x \in \Re^m_+$ such that $\sum_{\iota=1}^m \lambda_x^\iota = 1$ and $x = \sum_{\iota=1}^m \lambda_x^\iota v^\iota$. Furthermore, recall that the directional derivative is positively homogeneous, i.e., $rf'(x; d) = f'(x; rd)$ for all $r \geq 0$. Then, using the expression in (6) we get $0 \leq \sum_{\iota=1}^m \lambda_x^\iota f'(\bar{x}; v^\iota - \bar{x}) = \sum_{\iota=1}^m f'(\bar{x}; \lambda_x^\iota(v^\iota - \bar{x})) = \sum_{\iota=1}^m \min_{g \in \partial^{\mathsf{c}} f(\bar{x})} \langle g, \lambda_x^\iota(v^\iota - \bar{x})\rangle \leq \min_{g \in \partial^{\mathsf{c}} f(\bar{x})} \langle g, \sum_{\iota=1}^m \lambda_x^\iota v^\iota - \bar{x}\rangle$. The latter is equal to $\min_{g \in \partial^{\mathsf{c}} f(\bar{x})} \langle g, x - \bar{x}\rangle = f'(\bar{x}; x - \bar{x})$. As $x \in X$ is arbitrary, d-stationarity of $\bar{x}$ to (1) holds. ◀

This fact, already exploited in [3] for more structured functions, motivates the following escaping procedure to prevent Algorithm 1 from stopping at a C-stationary point that is not d-stationary: (i) let $\widetilde{x}$ be given by Algorithm 1; (ii) check if there exists a vertex v^ι of X such that $f'(\widetilde{x}; v^\iota - \widetilde{x}) < 0$ (such a derivative can be approximated by finite difference formulas); (iii) if such a vertex does not exist then stop ($\widetilde{x}$ is d-stationary), otherwise provide the descent direction $v^\iota - \widetilde{x}$ of f at $\widetilde{x}$ to the algorithm so that it can continue its iterative process. This approach is a roundabout for the more difficult task (already investigate in [4, 19, 25] using different strategies and convexity assumption): choose $\widetilde{g} \in \partial^{\mathsf{C}} f(\widetilde{x})$ such that the FW subproblem provides a direction of maximum descent for f at $\widetilde{x}$.

References

1 Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 26(1):115–129, 2015.

2 Amir Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2017.

3 Amir Beck and Nadav Hallak. On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM J. Optim.*, 30(1):56–79, 2020.

4 Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.*, 27(2):616–639, 2017.

5 Frank Clarke. *Optimisation and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1990.

6 Cyrille Combettes and Sebastian Pokutta. Boosting Frank–Wolfe by chasing gradients. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2111–2121. PMLR, 2020.

7 Aris Daniilidis and Jérôme Malick. Filling the gap between lower-C^1 and lower-C^2 functions. *J. Convex Anal.*, 12(2):315–329, 2005.

8 Welington de Oliveira. The ABC of DC programming. *Set-Valued Var. Anal.*, 28(4):679–706, 2020.

9 Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel. Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Probl.*, 36(1), 2020.

10 John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM J. Optim.*, 22(2):674–701, 2012.

11 Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 3(1-2):95–110, 1956.

12 Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435. PMLR, 2013.

13 Đ. Khuê Lê-Huu and Karteek Alahari. Regularized Frank–Wolfe for dense CRFs: Generalizing mean field and beyond. In *Advances in Neural Information Processing Systems*, volume 34, pages 1453–1467. Neural Information Processing Systems, 2021.

14 Hisashi Mine and Masao Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *J. Optim. Theory Appl.*, 33(1):9–23, 1981.

15 Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Math. Program.*, 171(1-2):311–330, 2018.

16 Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet K. Dokania, and Simon Lacoste-Julien. Minding the gaps for block Frank–Wolfe optimization of structured SVMs. In *Proceedings of the 33rd International Conference of Machine Learning*. ICML, 2016.

17 Jong-Shi Pang, Meisam Razaviyayn, and Alberth Alvarado. Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.*, 42(1):95–118, 2017.

18 Fabian Pedregosa, Geoffrey Negiar, Armin Askari, and Martin Jaggi. Linearly convergent Frank–Wolfe with backtracking line-search. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1–10. PMLR, 2020.

19 Sathya N. Ravi, Maxwell D. Collins, and Vikas Singh. A deterministic nonsmooth Frank–Wolfe algorithm with coreset guarantees. *INFORMS J. Optim.*, 1(2):120–142, 2019.

20 R. Tyrrell Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. In *Progress in Nondifferentiable Optimization*, IIASA Collaborative Proceedings Series, pages 125–144. International Institute of Applied Systems Analysis, 1982.

21 R. Tyrrell Rockafellar. Generalized subgradients in mathematical programming. In *Mathematical Programming The State of the Art, XIth International Symposium on Mathematical Programming, Bonn, Germany, August 23-27*, pages 23–27. Springer, 1982.

22 R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 2009.

23 Jonathan E. Spingarn. Submonotone subdifferentials of Lipschitz functions. *Trans. Am. Math. Soc.*, 264:77–89, 1981.

24 Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Projection efficient subgradient method and optimal nonsmooth Frank–Wolfe method. In *Advances in Neural Information Processing Systems 33*, pages 12211–12224. Curran Associates, Inc., 2020.

25 Douglas J. White. Extension of the Frank–Wolfe algorithm to concave nondifferentiable objective functions. *J. Optim. Theory Appl.*, 78(2):283–301, 1993.