Avoiding bad steps in Frank-Wolfe variants

Francesco Rinaldi* Damiano Zeffiro†

Abstract

The study of Frank-Wolfe (FW) variants is often complicated by the presence of different kinds of "good" and "bad" steps. In this article, we aim to simplify the convergence analysis of specific variants by getting rid of such a distinction between steps, and to improve existing rates by ensuring a non-trivial bound at each iteration.

In order to do this, we define the Short Step Chain (SSC) procedure, which skips gradient computations in consecutive short steps until proper conditions are satisfied. This algorithmic tool allows us to give a unified analysis and converge rates in the general smooth non convex setting, as well as a linear convergence rate under a Kurdyka-Łojasiewicz (KL) property. While the KL setting has been widely studied for proximal gradient type methods, to our knowledge, it has never been analyzed before for the Frank-Wolfe variants considered in the paper.

An angle condition, ensuring that the directions selected by the methods have the steepest slope possible up to a constant, is used to carry out our analysis. We prove that such a condition is satisfied, when considering minimization problems over a polytope, by the away step Frank-Wolfe (AFW), the pairwise Frank-Wolfe (PFW), and the Frank-Wolfe method with in face directions (FDFW).

Keywords: Nonconvex optimization, First-order optimization, Frank-Wolfe variants, Kurdyka-Łojasiewicz property.

Mathematics Subject Classification (2010): 46N10, 65K05, 90C06, 90C25, 90C30.

1 Introduction

The Frank-Wolfe method [22] and its variants (see, e.g., [23], [39] and references therein) provide a valid alternative to projected gradient approaches for the constrained optimization of a smooth objective $f : \mathbb{R}^n \to \mathbb{R}$, in settings where projecting on the feasible set may be unpractical. These methods have found many applications in sparse and structured optimization (see, e.g., [9], [23], [29], [31], [48] and references therein).

In this paper, we aim to overcome an annoying issue affecting the analysis of some FW variants, that is the presence of "bad iterations", i.e., iterations where we cannot show good progress. This happens when we are forced to take a short step along the search direction to guarantee feasibility of the iterate. The number of short steps typically needs to be upper bounded in the convergence analysis with "ad hoc" arguments (see, e.g., [23] and [39]). The main idea behind our method is to chain several short steps by skipping gradient updates until proper conditions are met.

1.1 Related work

FW variants. The main drawback of the classic FW algorithm is its slow $O(1/k)$ convergence rate for convex objectives. This rate is tight even for strongly convex objectives on polytopes, due to a well understood zig-zagging behaviour near optima on the boundary (see, e.g., [19] and [57]). The study of assumptions and variants leading to faster rates is a rapidly developing field.

*Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (rinaldi@math.unipd.it)
†Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (zeffiro@math.unipd.it)

Alternative or modified directions moving away from "bad" vertices or atoms have a long history, starting at least with the work of Wolfe [57] (see [37] and [39] for recent references). In addition to considering new directions, the works [16] and [17] propose strategies to skip the linear minimization oracle (LMO) computation from time to time by caching linear minimizers, while the recent work [37] for optimization on polytopes applies recursively a FW variant to smaller polytopes. However, to our knowledge, no strategy to avoid short steps has been discussed in these previous works.
For smooth strongly convex objectives, the convergence rates of many of these "improved directions" FW variants is linear on polytopes (see, e.g., [8] and [39]). Furthermore, in [35] it was proved that convergence rate of an AFW variant is adaptive to Hölderian error bound conditions interpolating between the general convex case and the strongly convex one.
A different approach, adopted in the general smooth convex setting, is to use FW variants to approximate projections (see, e.g., [20], [28] and [40]). To our knowledge, however, this approach always leads to a sublinear $O(1/\varepsilon)$ LMO complexity. Outside the projection free setting, in [46] a procedure making multiple steps without updating the gradient (in a fashion similar to our SSC) is defined, and it is claimed that the approach traces the piecewise linear projection curve on polytopes, thus leading to the same linear convergence rate of the standard projected gradient method in the strongly convex setting. In the non convex setting, for the classic FW algorithm a convergence rate of $O(1/\sqrt{k})$ was proved in [38] and then extended to other variants in [15] and [52].

KL property. The KL property (see, e.g., [4], [11] and [12]) has been extensively applied to compute the convergence rates of proximal subgradient type methods (see, e.g., [4], [5], [13], [56] and [58]). Furthermore, for convex objectives, it has been proved that Hölderian error bound conditions are a particular case of this property [13]. However, we are not aware of previous applications to the Frank-Wolfe variants under study in this paper.

Angle condition. The analysis of unconstrained descent methods often relies on some version of an angle condition, imposing an upper bound on the angle between the negative gradient and the descent direction selected by the method (see, e.g., [1], [25] and [59]). However, due to the presence of short steps and full FW steps, these analyses do not extend to our setting in a straightforward way.
In Section 3, we present an angle condition for optimization over a convex set. While to our knowledge this extension is novel for first order optimization methods, analogous conditions can be found in the context of direct search methods for linearly constrained derivative free optimization (see, e.g., [36] and [42]), imposed on the smallest angle between the negative gradient and a search direction. Finally, we remark that our condition was somehow used, but not stated explicitly, in [8] and [39] within the context of smooth strongly convex optimization over polytopes.

1.2 Contributions

Our main contributions are twofold:

- We formulate an angle condition for projection free methods, and prove that it leads to linear convergence in the number of "good steps" for non convex objectives satisfying a KL inequality (see Proposition 3.2). We show that this condition applies to the away step Frank-Wolfe (AFW), the pairwise Frank-Wolfe (PFW) and the FW method with in face directions (FDFW) (see, e.g., [23], [27] and [39]) on polytopes.

- We define the SSC procedure, which can be applied to all the FW variants listed in the first point, and show that it gets improvements on known rates (see Table 1 in Section 4). In particular, we prove that it leads to global linear convergence rates with no bad steps (see Theorem 4.1 and Corollary 4.1) under the KL inequality and the angle condition. This, to our knowledge, is the first (bad step free) linear convergence rate for FW variants under the KL inequality. In the general smooth non convex case, we further prove, under the angle condition, a $O(1/\sqrt{k})$ convergence rate with respect to a specific measure of non-stationarity for the iterates, that is the projection of the

negative gradient on the convex cone of feasible directions (see Theorem 4.2, Corollary 4.2 and Remark 4.3).

While here we apply our framework only to the AFW, the PFW, and the FDFW on polytopes, we remark that our results hold for projection free methods on generic convex sets. In an extended version of this paper [54] we show applications on convex sets with smooth boundary for FW variants and methods using orthographic retractions (see also [2], [6], [41] and references therein).
The reasons why eliminating bad steps truly makes a difference in our context are the following:

- it rules out impractical convergence rates due to a large number of bad steps. An interesting example is given by the rate guarantee reported in [39] for the pairwise Frank-Wolfe (PFW) variant on the $N-1$ dimensional simplex. This guarantee is indeed more loose than for the other variants, because there is no satisfactory bound on the number of such problematic steps (there is a best known bound of $3N!$ bad steps for each good step);
- it eliminates the dependence of the convergence rates on the support of the starting point (see, e.g., [30] and [37]). This dependence can significantly affect the performance of FW variants on smooth non convex optimization problems [21].

Finally, while beyond the scope of this paper, we mention that bad steps lead to a slow active set identification for the AFW, when compared to the "one shot" identification property characterizing proximal gradient methods and active set strategies (see [21], [47] and references therein). More precisely, analyses in recent works ([14], [15] and [24]) show that a number of bad steps equal to the number of "wrong" atoms is performed by the method in a sufficiently small neighborhood of a solution to identify its support.

1.3 Paper structure

The structure of the paper is as follows. In Section 2, we define some notation and state some preliminary results from convex analysis. In Section 3, we introduce the angle condition for first-order projection free methods, show examples of FW variants satisfying the condition and prove linear convergence in the number of good steps. We define the SSC procedure in Section 4, where we also state the main convergence results. The missing proofs can be found in the appendix.

2 Notation and preliminaries

We consider the following constrained optimization problem:

$$\min\{f(x) \mid x \in \Omega\}\,. \tag{2.1}$$

In the rest of the article Ω is a compact and convex set and $f \in C^1(\Omega)$ with L-Lipschitz gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \text{ for all } x, y \in \Omega\,.$$

We define D as the diameter of Ω, $\hat{c} = c/\|c\|$ for $c \in \mathbb{R}^n/\{0\}$ and $\hat{c} = 0$ for $c = 0$. For sequences we write $\{x_k\}$ instead of $\{x_k\}_{k\in I}$ when I is clear from the context, with $[j:i] = \{j, j+1, ..., i-1, i\}$.
For subsets C, D of $\mathbb{R}^n$ we define $\operatorname{dist}(C,D)$ as

$$\operatorname{dist}(C,D) = \inf\{\|y-z\| \mid z \in C,\ y \in D\}\,,$$

$B_R(C)$ as the neighborhood $\{x \in \mathbb{R}^n \mid \operatorname{dist}(C,x) < R\}$ of C of radius R and in particular $B_R(x)$ as the open euclidean ball of radius R and center x. When C is closed and convex we define as $\pi(C,\cdot)$ the projection on C. If C is a cone then we denote with C^* its polar.

We now state some elementary properties related to the tangent and the normal cones, where for $\bar{x} \in \Omega$ we denote with $T_\Omega(\bar{x})$ and $N_\Omega(\bar{x})$ the tangent and the normal cone to Ω in $\bar{x}$ respectively. The next proposition (from [55], Theorem 6.9) characterizes these cones for closed convex subsets of $\mathbb{R}^n$.

Proposition 2.1. *Let Ω be a closed convex set. For every point $\bar{x} \in \Omega$ we have*

$$\begin{aligned} T_\Omega(\bar{x}) &= \mathrm{cl}\{w \mid \exists \lambda > 0 \text{ with } \bar{x} + \lambda w \in \Omega\}, \\ \mathrm{int}(T_\Omega(\bar{x})) &= \{w \mid \exists \lambda > 0 \text{ with } \bar{x} + \lambda w \in \ \mathrm{int}(\Omega)\}, \\ N_\Omega(\bar{x}) &= T_\Omega(\bar{x})^* = \{v \in \mathbb{R}^n \mid (v, y - \bar{x}) \leq 0 \ \forall \ y \in \Omega\}. \end{aligned}$$

We have the following formula connecting the supremum of a linear function "slope" along feasible directions to the tangent and the normal cone:

Proposition 2.2. *If Ω is a closed convex subset of $\mathbb{R}^n$, $\bar{x} \in \Omega$ then for every $g \in \mathbb{R}^n$*

$$\max\left\{0, \sup_{h \in \Omega \setminus \{\bar{x}\}} \left(g, \frac{h - \bar{x}}{\|h - \bar{x}\|}\right)\right\} = \mathrm{dist}(N_\Omega(\bar{x}), g) = \|\pi(T_\Omega(\bar{x}), g)\|.$$

This property is a consequence of the Moreau-Yosida decomposition [55] and we refer the reader to the Appendix for a detailed proof. On polytopes, a geometric interpretation is that the smallest angle between g and a descent direction d feasible in $\bar{x}$ is achieved for $d = \pi(T_\Omega(\bar{x}), g)$.
In the rest of the article to simplify notations we often use $\pi_{\bar{x}}(g)$ as a shorthand for $\|\pi(T_\Omega(\bar{x}), g)\|$. Then, by Proposition 2.2, first order stationarity conditions in $\bar{x}$ for the gradient $-g$ become equivalent to $\pi_{\bar{x}}(g) = 0$.
In the computation of the convergence rates, we often assume

$$\pi_x(-\nabla f(x)) \geq \sqrt{2\mu}(f(x) - f(x^*))^{\frac{1}{2}} \tag{2.2}$$

for a fixed stationary point x^* and for every feasible point x of Problem (2.1). We refer the reader to the extended version [54] of this article for a study of convergence rates under a local condition (see Remark 3.2) as well as a more general inequality, interpolating between (2.2) and the generic non convex case. Let now i_Ω be the indicator function of Ω so that $i_\Omega(x) = 0$ in Ω and $i_\Omega(x) = +\infty$ otherwise. It can easily be seen that (2.2) is a special case of the KL inequality (see, e.g., [4], [5] and [13]) with exponent $\frac{1}{2}$

$$\mathrm{dist}(0, \partial f_\Omega(x)) \geq \sqrt{2\mu}(f_\Omega(x) - f_\Omega(x^*))^{\frac{1}{2}} \tag{2.3}$$

for $f_\Omega = f + i_\Omega$, using that

$$\pi_x(-\nabla f(x)) = \mathrm{dist}(-\nabla f(x), N_\Omega(x)) = \mathrm{dist}(0, \partial(f + i_\Omega)(x)), \tag{2.4}$$

with the last equality following by Proposition 2.2. Moreover, given that $\|\nabla f(x)\| \geq \pi_x(-\nabla f(x))$, condition (2.2) is weaker than the classic Polyak-Łojasiewicz inequality $\|\nabla f(x)\|^2 \geq \sqrt{2\mu}(f(x) - f(x^*))^{\frac{1}{2}}$ (from [44] and [51]), which in turn is implied by $\mu-$ strong convexity (see, e.g., [34]). Finally, condition (2.2) is locally implied by the Luo Tseng error bound [45] under some mild separability conditions for stationary points (see [43, Theorem 4.1]). This error bound is known to hold in a variety of convex and non convex settings (see references in [43]).

3 An angle condition

Let $\mathcal{A}$ be a first-order optimization method defined for smooth functions on a closed subset Ω of $\mathbb{R}^n$. We assume that given first-order information $(x_k, \nabla f(x_k))$ the method always selects x_{k+1} along a feasible descent direction, so that for $(x, g) \in \Omega \times \mathbb{R}^n$ we can define

$$\mathcal{A}(x, g) \subset T_\Omega(x) \cap \{y \in \mathbb{R}^n \mid \langle g, y \rangle > 0\} \cup \{0\}$$

as the possible descent directions selected by $\mathcal{A}$ when $x = x_k$, $g = -\nabla f(x_k)$ for some k (see Algorithm 1). When x is first-order stationary, we set $\mathcal{A}(x, g) = \{0\}$, otherwise we always assume $0 \notin \mathcal{A}(x, g) \neq \emptyset$.

Algorithm 1: First-order method

Initialization. $x_0 \in \Omega$, $k := 0$.
1. If x_k is stationary, then STOP
2. select a descent direction $d_k \in \mathcal{A}(x_k, -\nabla f(x_k))$
3. set $x_{k+1} = x_k + \alpha_k d_k$ for some stepsize $\alpha_k \in [0, \alpha_k^{\max}]$
4. set $k := k+1$, go to Step 1.

We want to formulate an angle condition for the descent directions selected by $\mathcal{A}$, with respect to the infimum of the angles achieved with feasible descent directions. In order to do that, we define the directional slope lower bound as

$$\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = \inf_{d \in \mathcal{A}(x,g)} \frac{\langle g, d \rangle}{\pi_x(g)\|d\|}$$

if $0 \notin \mathcal{A}(x,g)$. Otherwise x is stationary for $-g$, $\pi_x(g) = 0$ and we set $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = 1$. Then with this definition it immediately follows $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) \leq 1$ by Proposition 2.2. Notice also that when $x \in \mathrm{int}(\Omega)$ then $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g)$ is simply a lower bound on $\cos(\theta)$ with θ the angle between g and a descent direction d, and thus imposing $\mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) \geq \tau$ we retrieve the angle condition [1, equation (20)].
Given a subset P of Ω we can finally define the slope lower bound

$$\mathrm{SB}_{\mathcal{A}}(\Omega, P) = \inf_{\substack{g \in \mathbb{R}^n \\ x \in P}} \mathrm{DSB}_{\mathcal{A}}(\Omega, x, g) = \inf_{\substack{g : \pi_x(g) \neq 0 \\ x \in P}} \mathrm{DSB}_{\mathcal{A}}(\Omega, x, g)\,.$$

For simplicity if $P = \Omega$ we write $\mathrm{SB}_{\mathcal{A}}(\Omega)$ instead of $\mathrm{SB}_{\mathcal{A}}(\Omega, \Omega)$.

We now show a few examples of Frank-Wolfe variants satisfying the following *angle condition*

$$\mathrm{SB}_{\mathcal{A}}(\Omega) = \tau > 0, \tag{3.1}$$

i.e. cases where the slope lower bound is strictly greater than 0.

3.1 Frank-Wolfe variants over polytopes and the angle condition

We now consider the AFW, PFW and FDFW and show that the angle condition is satisfied when Ω is a polytope. The AFW and PFW depend on a set of "elementary atoms" A such that $\Omega = \mathrm{conv}(A)$. Given A, for a base point $x \in \Omega$ we can define

$$S_x = \{S \subset A \mid x \text{ is a proper convex combination of all the elements in } S\}\,,$$

the family of possible active sets for x. In the rest of the article A is always clear from the context and for simplicity we write PFW, AFW instead of PFW_A, AFW_A. For $x \in \Omega$, $S \in S_x$, d^{PFW} is a PFW direction with respect to the active set S and gradient $-g$ iff

$$d^{\mathrm{PFW}} = s - q \text{ with } s \in \mathrm{argmax}_{s \in \Omega} \langle s, g \rangle \text{ and } q \in \mathrm{argmin}_{q \in S} \langle q, g \rangle\,.$$

Similarly, given $x \in \Omega$, $S \in S_x$, d^{AFW} is an AFW direction with respect to the active set S and gradient $-g$ iff

$$d^{\mathrm{AFW}} \in \mathrm{argmax}\{\langle g, d \rangle \mid d \in \{d^{\mathrm{FW}}, d^{AS}\}\}\,, \tag{3.2}$$

where d^{FW} is a classic Frank-Wolfe direction

$$d^{\mathrm{FW}} = s - x \text{ with } s \in \mathrm{argmax}_{s \in \Omega} \langle s, g \rangle\,, \tag{3.3}$$

and d^{AS} is the away direction

$$d^{\mathrm{AS}} = x - q \text{ with } q \in \operatorname{argmin}_{q \in S} \langle q, g \rangle . \tag{3.4}$$

The FDFW from [23], [27] (sometimes referred to as Decomposition invariant Conditional Gradient (DiCG) when applied to polytopes [7]) relies only on the current point x and the current gradient $-g$ to choose a descent direction and, unlike the AFW and the PFW, does not need to keep track of the active set.

The in face direction is defined as

$$d^F = x_k - x_F \text{ with } x_F \in \operatorname{argmin}\{\langle g, y \rangle \mid y \in \mathcal{F}(x)\}$$

for $\mathcal{F}(x)$ the minimal face of Ω containing x. The selection criterion is then analogous to the one used by the AFW:

$$d^{\mathrm{FD}} \in \operatorname{argmax}\{\langle g, d \rangle \mid d \in \{d^F, d^{\mathrm{FW}}\}\} . \tag{3.5}$$

We write $\mathrm{SB}_{\mathrm{FD}}, \mathrm{DSB}_{\mathrm{FD}}$ instead of $\mathrm{SB}_{\mathrm{FDFW}}, \mathrm{DSB}_{\mathrm{FDFW}}$ in the rest of the paper. When Ω is a polytope and $|A| < \infty$, the angle condition holds for the directions and the related FW variants we introduced. Before stating a lower bound for $\mathrm{SB}_{\mathcal{A}}(\Omega)$ in this setting we need to recall the definition of pyramidal width $\mathrm{PWidth}(A)$ as it was given in [39]. We refer the reader to [49] for a discussion of various properties of this parameter.

For a given $g \in \mathbb{R}^n \setminus \{0\}$ the pyramidal directional width is defined as

$$\mathrm{PdirW}(A, g, x) = \min_{S \in S_x} \max_{\substack{a \in A \\ s \in S}} \langle \frac{g}{\|g\|}, a - s \rangle , \tag{3.6}$$

and the pyramidal width is defined as

$$\mathrm{PWidth}(A) = \min_{\substack{F \in \mathrm{faces}(\mathrm{conv}(A)), x \in F \\ g \in \mathrm{cone}(F - x) \setminus \{0\}}} \mathrm{PdirW}(F \cap A, g, x) .$$

Here we use one key property of $\mathrm{PWidth}(A)$ which relates it to the slope along the PFW direction. We have the following lower bound (see [39, equation (12)])

$$\frac{\langle g, d^{\mathrm{PFW}} \rangle}{\langle g, \hat{e} \rangle} \geq \mathrm{PWidth}(A) > 0 , \tag{3.7}$$

where e is any direction in $\Omega - \{x\} = T_\Omega(x)$ which is also a descent direction[1] for $-g$. Another relevant property is that by (3.6) when Ω is fixed $\mathrm{PWidth}(A)$ is monotone decreasing in A, so that if $V(\Omega)$ is the set of vertexes of Ω we always have $\mathrm{PWidth}(A) \leq \mathrm{PWidth}(V(\Omega))$.

Proposition 3.1. $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \tau_p := \frac{\mathrm{PWidth}(A)}{D}, \mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\tau_p}{2}, \mathrm{SB}_{\mathrm{FD}}(\Omega) \geq \frac{\tau_v}{2} := \frac{\mathrm{PWidth}(V(\Omega))}{2D}$.

Proof. Let g be such that $\pi_x(g) \neq 0$. Then there exists descent directions for $-g$ feasible for Ω from x, and

$$0 < \max_{e \in \Omega - \{x\}} \langle g, \hat{e} \rangle ,$$

so that

$$\min_{\substack{e \in \Omega - \{x\}, \\ \langle g, \hat{e} \rangle > 0}} \frac{1}{\langle g, \hat{e} \rangle} = \frac{1}{\max_{e \in \Omega - \{x\}} \langle g, \hat{e} \rangle} = \frac{1}{\sup_{h \in \Omega \setminus \{\bar{x}\}} \left\langle g, \frac{h - \bar{x}}{\|h - \bar{x}\|} \right\rangle} = \frac{1}{\|\pi(T_\Omega(x), g)\|} , \tag{3.8}$$

[1] In [39] the direction e is defined as a possible error direction $e = x^* - x$, where x^* is an optimal point of a convex objective f with $\nabla f(x) = -g$. However, this definition is equivalent to ours. Indeed if $e = x^* - x$ then by convexity it must be a feasible descent direction for $-g$. Conversely, every feasible descent direction is always an error direction as defined above for some choice of f, i.e. consider $f(y) = \frac{1}{2}\langle y - x^*, Q(y - x^*) \rangle$ with Q positive definite such that $\nabla f(x) = Q(x - x^*) = -g$.

where we used Proposition 2.2 in the last equality. Thanks to (3.8) taking the min on all the feasible descent directions e in the LHS of (3.7) we obtain

$$\frac{\langle g, d^{\mathrm{PFW}}\rangle}{\|\pi(T_\Omega(x), g)\|} \geq \mathrm{PWidth}(A)\,.$$

We now have

$$\mathrm{DSB}_{\mathrm{PFW}}(\Omega, x, g) = \inf_{d^{\mathrm{PFW}} \in \mathrm{PFW}(x,g)} \frac{\langle g, d^{\mathrm{PFW}}\rangle}{\|d^{\mathrm{PFW}}\|\|\pi(T_\Omega(x), g)\|} \geq \frac{\langle g, d^{\mathrm{PFW}}\rangle}{D\|\pi(T_\Omega(x), g)\|} \geq \frac{\mathrm{PWidth}(A)}{D}\,.$$

Hence $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \frac{\mathrm{PWidth}(A)}{D}$ follows by taking the inf on the LHS for $x \in \Omega$ and g such that $\pi_x(g) \neq 0$ in (3.1). The inequality $\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\mathrm{PWidth}(A)}{2D}$ is a corollary since

$$\langle g, d^{\mathrm{AFW}}\rangle \geq \frac{1}{2}\langle g, d^{\mathrm{PFW}}\rangle\,,$$

as it follows immediately from the definitions (see also [39, equation (6)]).

For the FDFW, first observe that for any $S \in S_x$:

$$\begin{aligned}\langle g, d^{\mathrm{FW}} + d^F\rangle = \max\{\langle g, s-q\rangle \mid s \in \Omega,\ q \in \mathcal{F}(x)\} &\geq \max\{\langle g, s-q\rangle \mid s \in \Omega,\ q \in S\} \\ &= \langle g, d^{\mathrm{PFW}}\rangle\,,\end{aligned} \tag{3.9}$$

where we used $\mathcal{F}(x) \supset S$ in the inequality and the last equality follows immediately from the definition of d^{PFW}. We then have

$$\mathrm{DSB}_{\mathrm{FD}}(\Omega, x, g) = \inf_{d^{\mathrm{FD}} \in \mathrm{FDFW}(x,g)} \frac{\langle g, d^{\mathrm{FD}}\rangle}{\pi_x(g)\|d^{\mathrm{FD}}\|} \geq \frac{\langle g, d^{\mathrm{FW}} + d^F\rangle}{2D\pi_x(g)} \geq \frac{\langle g, d^{\mathrm{PFW}}\rangle}{2D\pi_x(g)} \geq \frac{\mathrm{PWidth}(A)}{2D}\,, \tag{3.10}$$

where the first inequality follows from the choice criterion (3.5) together with $\|d^{\mathrm{FD}}\| \leq 2D$, we used (3.9) in the second inequality, and (3.7) in the third. The thesis follows by taking the inf on x and g in the LHS and the sup on A in the RHS. □

Remark 3.1. *Results analogous to the ones in Proposition 3.1 can be proven relatively to the vertex facial distance* $\mathrm{vf}(\Omega)$ *from [8]. More precisely, assuming* $A = V(\Omega)$ *and that the AFW and the PFW keep active sets of size at most* $\bar{s}$*, we have* $\mathrm{SB}_{\mathrm{PFW}}(\Omega) \geq \frac{\mathrm{vf}(\Omega)}{\bar{s}D}$*,* $\mathrm{SB}_{\mathrm{AFW}}(\Omega) \geq \frac{\mathrm{vf}(\Omega)}{2\bar{s}D}$ *as a consequence of [8, Lemma 3.1]. Furthermore, for the FDFW we have* $\mathrm{SB}_{\mathrm{FD}}(\Omega, \Omega_{\bar{s}}) \geq \frac{\mathrm{vf}(\Omega)}{2\bar{s}D}$*, with* $x \in \Omega_{\bar{s}} \subset \Omega$ *iff there exists* $S \in S_x$ *such that* $|S| \leq \bar{s}$*.*

3.2 Linear convergence for good steps under the angle condition

Consider now a method following the scheme described by Algorithm 1 and with stepsize given by

$$\alpha_k = \min\left(\bar{\alpha}_k, \alpha_k^{\max}\right)\,, \tag{3.11}$$

where

$$\bar{\alpha}_k = \frac{\langle -\nabla f(x_k), d_k\rangle}{L\|d_k\|^2}. \tag{3.12}$$

We notice that $\bar{\alpha}_k$ in (3.12) is a standard stepsize, often used in numerical tests with a properly tuned estimate for L (see, e.g., [50]).

Let $\{x_k\}$ be a sequence generated by Algorithm 1, with a method $\mathcal{A}$ satisfying the angle condition (3.1). We say that the algorithm performs a *full FW step* if

$$x_{k+1} \in \mathrm{argmin}_{x \in \Omega}\langle \nabla f(x_k), x\rangle\,. \tag{3.13}$$

We now prove a general linear convergence rate in the number of *good steps*, i.e. the steps satisfying $\alpha_k = \bar{\alpha}_k$ or (3.13), under the assumption that the method $\mathcal{A}$ satisfies the angle condition (3.1), and the objective function f in Problem (2.1) satisfies the KL inequality (2.2).

Proposition 3.2. *Let us assume that $\mathcal{A}$ satisfies the angle condition* (3.1)*, and the objective function f in Problem* (2.1) *satisfies condition* (2.2)*.*

- *If $\alpha_k = \bar{\alpha}_k$ then*
$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\tau^2\right)(f(x_k) - f(x^*)). \tag{3.14}$$
- *If the step k is a full FW step then*
$$f(x_{k+1}) - f(x^*) \leq \left(1 + \frac{\mu}{L}\right)^{-1}(f(x_k) - f(x^*)). \tag{3.15}$$

Proof. Let $p_k = \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_{k+1}))\|$ and $\tilde{p}_k = \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_k))\|$. We have

$$\begin{aligned} |p_k - \tilde{p}_k| &= |\|\pi(T_\Omega(x_{k+1}), -\nabla f(x_{k+1}))\| - \|\pi(T_\Omega(x_{k+1}), -\nabla f(x_k))\|| \\ &\leq \|-\nabla f(x_{k+1}) + \nabla f(x_k)\| \leq L\|x_{k+1} - x_k\|, \end{aligned} \tag{3.16}$$

where we used the 1-Lipschitzianity of projections in the first inequality.
Moreover, by the standard descent lemma [10, Proposition 6.1.2],

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \leq f(x_k) + \alpha_k \langle \nabla f(x_k), d_k\rangle + \alpha_k^2 \frac{L}{2}\|d_k\|^2, \tag{3.17}$$

and in particular

$$f(x_k) - f(x_{k+1}) \geq -\alpha_k \langle \nabla f(x_k), d_k\rangle - \alpha_k^2 \frac{L}{2}\|d_k\|^2 \geq \frac{L}{2}\alpha_k^2\|d_k\|^2 = \frac{L}{2}\|x_{k+1} - x_k\|^2, \tag{3.18}$$

where we used $\alpha_k \leq \bar{\alpha}_k$ in the last inequality.
If $\alpha_k = \bar{\alpha}_k$ then

$$\begin{aligned} f(x_{k+1}) =& f(x_k + \bar{\alpha}_k d_k) \leq f(x_k) - \frac{1}{2L}\left(\frac{\langle \nabla f(x_k), d_k\rangle}{\|d_k\|}\right)^2 \leq f(x_k) - \frac{\tau^2}{2L}p_{k-1}^2 \\ \leq& f(x_k) - \frac{\mu\tau^2}{L}(f(x_k) - f(x^*)), \end{aligned} \tag{3.19}$$

where we used (3.17) in the first inequality, $\mathrm{SB}_{\mathcal{A}}^f(\Omega) = \tau$ in the second one, and condition (2.2) in the third one.
If the step k is a full FW step then

$$f(x_{k+1}) - f(x^*) \leq \frac{p_k^2}{2\mu} \leq \frac{(\tilde{p}_k + L\|x_{k+1} - x_k\|)^2}{2\mu} = \frac{L^2}{2\mu}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\mu}(f(x_k) - f(x_{k+1})), \tag{3.20}$$

where we used (2.2) in the first inequality, (3.16) in the second, $\tilde{p}_k = 0$ by stationarity conditions, and (3.18) in the last inequality. Then (3.18) and (3.15) follow by rearranging (3.19) and (3.20) respectively. □

Remark 3.2. *The assumption that the objective satisfies condition* (2.2) *can be replaced by a (local) KL property together with a lower bound on the values of the objective and some mild conditions on the starting point, as it can be proved adapting well known techniques from [4], [5]. We refer the reader to the extended version of this article [54] for additional details.*

For the three FW variants described before we can now give a linear convergence rate in the number of good steps. We refer the reader to Table 2 for bounds on this number.

Corollary 3.1. *Let us assume that the objective function f satisfies condition* (2.2) *and* $\Omega = \mathrm{conv}(A)$ *with* $|A| < +\infty$ *in Problem* (2.1)*. Then the AFW, the PFW and the FDFW converge at a rate*

$$f(x_k) - f(x^*) \leq \bar{q}_{gs}^{\bar{\gamma}(k)}(f(x_0) - f(x^*)), \tag{3.21}$$

with $\bar{\gamma}(k)$ *the number of good steps among the first* k *steps,*

$$\bar{q}_{gs} = \max\left(1 - \frac{\mu}{L}\left(\frac{\mathrm{PWidth}(A)}{2D}\right)^2, \left(1 + \frac{\mu}{L}\right)^{-1}\right) \tag{3.22}$$

for the AFW,

$$\bar{q}_{gs} = 1 - \frac{\mu}{L}\left(\frac{\mathrm{PWidth}(A)}{D}\right)^2 \tag{3.23}$$

for the PFW, and

$$\bar{q}_{gs} = \max\left(1 - \frac{\mu}{L}\left(\frac{\mathrm{PWidth}(V(\Omega))}{2D}\right)^2, \left(1 + \frac{\mu}{L}\right)^{-1}\right) \tag{3.24}$$

for the FDFW.

Proof. For the AFW and the FDFW the rates (3.22) and (3.24) follows directly from (3.14) and (3.15) together with the bound on τ given in Proposition 3.1. Since the PFW never performs full FW steps, its rate (3.23) follow directly from (3.14) together with the bound on τ given in Proposition 3.1. Finally, given that by (3.18) the sequence $\{f(x_k)\}$ is decreasing, (3.21) follows from the rate for good steps by induction. □

4 First order projection free methods with SSC procedure

We introduce here the SSC procedure, and prove convergence rates both under the KL inequality (2.2) and in the generic non convex case.

4.1 The SSC procedure

The SSC procedure chains consecutive short steps, thus skipping updates for the gradient (and possibly for related information, like linear minimizers), until proper stopping conditions are met. Such a procedure, whose detailed scheme is given in Algorithm 3, can be easily embedded in a first-order approach (see Algorithm 2).

Algorithm 2: First-order method with SSC

Initialization. $x_0 \in \Omega$, $k = 0$.
1. **while** x_k is not stationary:
2. $\quad g = -\nabla f(x_k)$
3. $\quad x_{k+1} = \mathrm{SSC}(x_k, g)$
5. $\quad k = k + 1$.

Given that the gradient $-g$ is constant during the SSC, this procedure is an application of $\mathcal{A}$ for the minimization of the linearized objective $f_g(z) = \langle -g, z - \bar{x}\rangle + f(\bar{x})$ with peculiar stepsizes and stopping criterion. More specifically, after a stationarity check (Phase I), the stepsize α_j is computed by taking the minimum between the maximal stepsize $\alpha_{\max}^{(j)}$ (which we always assume to be greater than 0) and

Algorithm 3: SSC$(\bar{x}, g)$

Initialization. $y_0 = \bar{x}$, $j = 0$.
Phase I
1. select $d_j \in \mathcal{A}(y_j, g)$, $\alpha_{\max}^{(j)} \in \alpha_{\max}(y_j, d_j)$
2. **if** $d_j = 0$ **then:**
3. **return** y_j

Phase II
4. compute β_j with (4.2)
5. let $\alpha_j = \min(\alpha_{\max}^{(j)}, \beta_j)$
6. $y_{j+1} = y_j + \alpha_j d_j$
7. **if** $\alpha_j = \beta_j$ **then**:
8. **return** y_{j+1}
9. $j = j + 1$, go to Step 1.

an auxiliary stepsize β_j. Notice that the SSC always terminates after a FW step (see equation (3.3)), with $\alpha_j = \beta_j$ or with y_{j+1} global minimizer of f_g.

The auxiliary step size β_j is defined as the maximal feasible stepsize for the trust region

$$\Omega_j = \bar{B}_{\|g\|/2L}(\bar{x} + \frac{g}{2L}) \cap \bar{B}_{\langle g, \hat{d}_j \rangle / L}(\bar{x}) \tag{4.1}$$

when $y_j \in \Omega_j$, otherwise the method stops returning y_j. Summarizing,

$$\beta_j = \begin{cases} 0 & \text{if } y_j \notin \Omega_j\,, \\ \beta_{\max}(\Omega_j, y_j, d_j) & \text{if } y_j \in \Omega_j\,, \end{cases} \tag{4.2}$$

where $\beta_{\max}(\Omega_j, y_j, d_j) = \max\{\beta \in \mathbb{R}_{\geq 0} \mid y_j + \beta d_j \in \Omega_j\}$ is the maximal feasible stepsize in the direction d_j starting from y_j with respect to Ω_j. Since Ω_j is the intersection of two balls there is a simple closed form expression for β_j. In particular, using that $y_0 = \bar{x}$, if $d_0 \neq 0$ we have

$$\beta_0 = \frac{\langle g, \hat{d}_0 \rangle}{L\|d_0\|}\,,$$

which corresponds to (3.11) in the non maximal case, and where $\beta_0 > 0$ since $d_0 \neq 0$ is by assumption a descent direction for $-g$.

Employing the trust region Ω_j in the definition of β_j guarantees the sufficient decrease condition

$$f(y_j) \leq f(x_k) - \frac{L}{2}\|x_k - y_j\|^2 \tag{4.3}$$

and monotonicity of the true objective f during the SSC.

To see why (4.3) holds, notice that the second ball $\bar{B} = \bar{B}_{\|g\|/2L}(x_k + \frac{g}{2L})$ appearing in the definition of Ω_j does not depend on j, so that since $y_0 \in \bar{B}$ we have $y_j \in \bar{B}$ for every $j \in [0:T]$, with T maximal iteration index of the SSC. This is enough to obtain (4.3) because for every $z \in \bar{B}$ we have

$$f(z) \leq f(\bar{x}) - \langle g, z - \bar{x} \rangle + \frac{L}{2}\|z - \bar{x}\|^2 \leq f(\bar{x}) - \frac{L}{2}\|\bar{x} - z\|^2\,, \tag{4.4}$$

where the first inequality is the standard descent lemma and the second follows from the definition of $\bar{B}$.

We prove that the true objective f is monotone decreasing in the next lemma.

Lemma 4.1. *Let us assume $y_j \in \bar{B}_{\langle g,\hat{d}_j\rangle/L}(\bar{x})$. Then for every $\beta \in [0,\beta_j]$ we have*

$$\frac{d}{d\beta} f(y_j + \beta d_j) \leq 0\,,$$

and thus in particular $f(y_j + \beta_j d_j) \leq f(y_j)$.

Proof. We have

$$\begin{aligned}
&\frac{d}{d\beta} f(y_j + \beta d_j) = \|d_j\| \langle \nabla f(y_j + \beta d_j), \hat{d}_j \rangle \\
=&\|d_j\| \langle (\nabla f(y_j + \beta d_j) + g) - g, \hat{d}_j \rangle = \|d_j\| (\langle \nabla f(y_j + \beta d_j) + r, \hat{d}_j \rangle - \langle g, \hat{d}_j \rangle) \\
\leq&\|d_j\| (L\|\bar{x} - y_j - \beta d_j\| - \langle g, \hat{d}_j \rangle) \leq 0\,,
\end{aligned}$$

where we used $g = -\nabla f(\bar{x})$ and the Lipschitzianity of ∇f in the first inequality and

$$y_j + \beta d_j \in \bar{B}_{\langle g,\hat{d}_j\rangle/L}(\bar{x})$$

in the second. □

The next result illustrates how the sequence $\{x_k\}$ generated by Algorithm 2 satisfies certain descent conditions. This is an adaptation to our setting of the ones used in the analysis of many proximal type gradient methods (see [4], [5], [13] and references therein). A subtle difference is the introduction of an "hidden sequence" $\{\tilde{x}_k\}$ to control the projection of the negative gradient on the tangent cone.

Proposition 4.1. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 2 and assume that*

- *the angle condition* (3.1) *holds;*
- *the SSC condition terminates in a finite number of steps.*

Then

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{2}\|x_k - x_{k+1}\|^2\,, \tag{4.5}$$

$$\|x_k - x_{k+1}\| \geq K\|\pi(T_\Omega(\tilde{x}_k), -\nabla f(\tilde{x}_k))\| \tag{4.6}$$

for some $\tilde{x}_k \in \{y_j\}_{j=0}^T$ such that $f(x_{k+1}) \leq f(\tilde{x}_k) \leq f(x_k) - \frac{L}{2}\|x_k - \tilde{x}_k\|^2$ and for $K = \tau/(L(1+\tau))$.

Remark 4.1. *From Proposition 4.1 under the KL inequality* (2.2) *it is easy to obtain the linear decrease equation (see Theorem 4.1)*

$$f(x_k) - f(x_{k+1}) \geq \frac{\mu\tau^2}{2L(1+\tau)^2}(f(x_{k+1}) - f(x^*))\,. \tag{4.7}$$

This improvement does not strictly depend from the choice of β_j given in (4.2)*. For instance, it is easy to check that we still have* (4.7) *if β_j is given by linesearch. In this particular case, if the method $\mathcal{A}$ is affine invariant then the SSC is too, and we have $f(x_k) - f(x_{k+1}) \geq q_{af}(f(x_k) - f(x^*))$ for some affine invariant constant q_{af}, with $q_{af} \geq \frac{\mu L K^2}{2}$ by* (4.7)*.*

4.2 SSC for Frank-Wolfe variants

In this section, we show how to apply our results to the PFW, the AFW and the FDFW on polytopes, i.e., we prove finite termination of the SSC procedure when one of these methods is considered in Algorithm 2.

Let $\{S^{(j)}\}$ be the sequence of active sets generated by the AFW and the PFW method in the SSC, with y_j proper convex combination of the elements in $S^{(j)}$. By controlling the cardinality of $S^{(j)}$ we can easily prove finite SSC termination for theses methods.

Proposition 4.2. *If no points are added to the active set beside the linear minimizer, the SSC with AFW (or PFW) terminates after at most $|S^{(0)}|$ steps.*

Proof. After a PFW step, the linear minimizer s is always in the active set. In particular, $S^{(j+1)} \subset S^{(j)} \cup \{s\}$. Now on the one hand, a linear minimizer added to the active set cannot be dropped from it, since the gradient never changes. On the other hand, after a maximal away or PFW step the element q_j corresponding to the away direction is dropped from the active set. Finally, the SSC terminates after a FW step, as anticipated in Section 4.1. Thus for both methods at most $|S^{(0)}|$ elements can be dropped from the active set, and the thesis follows. □

For the FDFW we assume that the maximal stepsize is given by feasibility conditions as in [23]:

$$\alpha_{\max}(x, d) = \max\{\alpha \in \mathbb{R}_{\geq 0} \mid x + \alpha d \in \Omega\}. \tag{4.8}$$

Then after a maximal in face step from y_j we have $\dim(\mathcal{F}(y_{j+1})) < \dim(\mathcal{F}(y_j))$ because y_{j+1} lies on the boundary of $\mathcal{F}(y_j)$. Whence there can only be a finite number of consecutive such steps, and given that after a FW step the SSC terminates, we have the following:

Proposition 4.3. *The FDFW on any compact and convex set Ω performs at most $\dim(\Omega) + 1$ consecutive maximal steps. In particular, for this method the SSC terminates after a finite number of iterations.*

4.3 Convergence rates

As a consequence of Proposition 4.1, we have linear convergence rates for the general algorithmic scheme reported in Algorithm 2 under the KL inequality (2.2), the angle condition (3.1), and finite termination of the SSC procedure.

Theorem 4.1. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 2 and assume that*

- *the objective function f satisfies condition* (2.2)*;*
- *the angle condition* (3.1) *holds;*
- *the SSC procedure always terminates in a finite number of steps.*

Then, for $q = \left(1 + \frac{\mu}{L}\frac{\tau^2}{(1+\tau)^2}\right)^{-1}$ we have $f(x_k) \to f(x^*)$, with

$$f(x_k) - f(x^*) \leq q^k (f(x_0) - f(x^*)), \tag{4.9}$$

and $x_k \to \tilde{x}^*$ with

$$\|x_k - \tilde{x}^*\| \leq \frac{\sqrt{2-2q}(f(x_0) - f(\tilde{x}^*))}{\sqrt{L}(1-\sqrt{q})} q^{\frac{k}{2}}, \tag{4.10}$$

for $\tilde{x}^*$ stationary point such that $f(\tilde{x}^*) = f(x^*)$.

Remark 4.2. *As for Proposition 3.2, we can replace condition* (2.2) *with a local KL property in this case (see Remark 3.2).*

We now give a corollary for Theorem 4.1 related to the FW variants described in Section 3.1.

Corollary 4.1. *Let us assume that the objective function f satisfies condition* (2.2) *and $\Omega = \mathrm{conv}(A)$ with $|A| < +\infty$ in Problem* (2.1)*. Then the sequence $\{x_k\}$ generated by Algorithm 2 with AFW (PFW or FDFW) in the SSC converges at the rates given by Theorem 4.1, with $\tau = \tau_p/2$ (τ_p or $\tau_v/2$, respectively).*

Proof. Finite termination of the SSC follows by Proposition 4.2 and Proposition 4.3, and the angle condition is satisfied by Proposition 3.1. Thus we have all the assumptions to apply Theorem 4.1. □

Algorithm	Article	Objective	$\gamma(k)$	I_b	q_{gs}	h_k/h_0 upper bound
AFW	[39]	SC	$k/2$	$\lvert S_0\rvert-1$	$1-\frac{\mu}{L}\frac{\tau_p^2}{4}$	$\left(1-\frac{\mu}{L}\frac{\tau_p^2}{4}\right)^{\frac{k}{2}}$
PFW	[39]	SC	$k/(3\lvert A\rvert!+1)$	-	$1-\frac{\mu}{L}\tau_p^2$	$\left(1-\frac{\mu}{L}\tau_p^2\right)^{\frac{k}{3\lvert A\rvert!+1}}$
FDFW[2]	[37]	SC	$k/(\Delta(\Omega)+1)$	$\dim(\mathcal{F}(x_0))$	$1-\frac{\mu}{L}\frac{\tau_v^2}{4}$	$\left(1-\frac{\mu}{L}\frac{\tau_v^2}{4}\right)^{\frac{k}{\Delta(\Omega)+1}}$
AFW + SSC	Ours	NC, KL	k	-	$\left(1+\frac{\mu}{L}\frac{\tau_p^2}{(2+\tau_p)^2}\right)^{-1}$	$\left(1+\frac{\mu}{L}\frac{\tau_p^2}{(2+\tau_p)^2}\right)^{-k}$
PFW + SSC	Ours	NC, KL	k	-	$\left(1+\frac{\mu}{L}\frac{\tau_p^2}{(1+\tau_p)^2}\right)^{-1}$	$\left(1+\frac{\mu}{L}\frac{\tau_p^2}{(1+\tau_p)^2}\right)^{-k}$
FDFW + SSC	Ours	NC, KL	k	-	$\left(1+\frac{\mu}{L}\frac{\tau_v^2}{(1+\tau_v)^2}\right)^{-1}$	$\left(1+\frac{\mu}{L}\frac{\tau_v^2}{(1+\tau_v)^2}\right)^{-k}$

Table 1: Comparison between the rates of the standard and SSC version of some FW variants for $\Omega=\operatorname{conv}(A)$ with $|A|<\infty$. SC = strongly convex, NC = non convex, KL = KL property. $\gamma(k)$: lower bound on the number of good steps after k steps, counting from the first good step. I_b: bound on the number of bad steps before the first good step. q_{gs}: rate in good steps. h_k/h_0 upper bound: worst case rate assuming no initial bad steps, equal to $q_{gs}^{\gamma(k)}$. $\Delta(\Omega)=$ maximum increase in face dimension $\mathcal{F}(x_{k+1})-\mathcal{F}(x_k)$ after a FW step. $S_0=$ active set for x_0.

For comparison, we now recall some well-known result related to global linear convergence rates for the FW variants under analysis.

Proposition 4.4. *Let us assume that the objective function f is $\mu-$strongly convex and $\Omega=\operatorname{conv}(A)$ with $|A|<+\infty$ in Problem (2.1). Let $\{x_k\}$ be a sequence generated by the AFW (PFW or FDFW), with stepsize given by linesearch. If the initial active set is $S_0=\{x_0\}$ for the AFW ($S_0=\{x_0\}$ for the PFW, $\dim(\mathcal{F}(x_0))=0$ for the FDFW), then*

$$f(x_k)-f^*\leq q_{gs}^{\gamma(k)}(f(x_0)-f^*)\,, \tag{4.11}$$

for $\gamma(k)$ and q_{gs} given in Table 1.

Proof. For the AFW and the PFW the result follows directly from [39, Theorem 1], with the exception of the good steps rate for the PFW, which can be obtained by applying the bound [39, Equation 10] in [39, Equation 5]. For the FDFW the result follows from [37, Theorem 1] (where the method is referred to as DiCG), by using the bound $\mu\mathrm{PWidth}(V(\Omega))^2$ on the geometric strong convexity constant implied by [39, Theorem 6]. □

For all the examples where an upper bound on $\tau_p=\frac{\mathrm{PWidth}(A)}{D}$ is known (see [49], [53] and references therein) when $\dim(\operatorname{conv}(A))\to\infty$ then $\tau_p\to 0$ and our rates for the SSC converge to the rates without SSC for good steps in Table 1. While we are not able to prove this limit in general, for all polytopes with dimension greater or equal to 2, except low dimensional simplices (see Example 4.1), we still have $\tau_p\leq\frac{1}{2}$ (because $\mathrm{PdirW}(A,g,x)+\mathrm{PdirW}(A,-g,x)\leq D$ for x in the relative interior of $\operatorname{conv}(A)$ and $\pm g$ feasible and orthogonal to $\operatorname{conv}(S)$ for some $S\in S_x$). Using this together with Example 4.1 for simplices, it is easy to check that the rates in Corollary 4.1 (SSC based FW variants) are strict improvements on the known worst case rates (standard FW variants) reported in Proposition 4.4, with a limited number of exceptions. These are the trivial one dimensional case and simplices with low dimension (≤ 4 for the PFW, and ≤ 8 for the AFW using the loose bounds in Example 4.1) combined with objectives having condition number μ/L sufficiently close to 1.

Example 4.1. *If $W(\mathrm{conv}(A))$ is the width of $\mathrm{conv}(A)$ (see [39, Section 3]) then it follows directly from the definition of* PWidth *that $W(\mathrm{conv}(A)) \geq \mathrm{PWidth}(A)$, with equality for $A = \{e_1, ..., e_n\}$ (see [39] and [49]). Let now $A = \{a_1, ..., a_n\}$ be a set of n affinely independent points in $\mathbb{R}^{n-1}$. We claim that, for $r_n = \sqrt{1 - \frac{1}{n}}$ circumradius of Δ_{n-1}*

$$\mathrm{PWidth}(A)/D \leq r_n^{-1} W(\Delta_{n-1}) = \begin{cases} 2r_n^{-1}\sqrt{\frac{1}{n}} & \text{for } n \text{ even}, \\ 2r_n^{-1}\sqrt{\frac{1}{n-1/n}} & \text{for } n \text{ odd}. \end{cases} \tag{4.12}$$

To see this, assume without loss of generality $D = 1$ and $0 \in \mathrm{int}(\Omega)$ for $\Omega = \mathrm{conv}(A)$. Then if $\hat{A} = \{\hat{a}_1, ..., \hat{a}_n\}$ we have $W(\mathrm{conv}(\hat{A})) \geq W(\mathrm{conv}(A))$. We can conclude

$$\frac{\mathrm{PWidth}(A)}{D} = \mathrm{PWidth}(A) \leq W(\mathrm{conv}(A)) \leq W(\mathrm{conv}(\hat{A})) \leq r_n^{-1} W(\Delta_{n-1}), \tag{4.13}$$

where in the last inequality we used that regular simplices maximize the width among simplices with fixed inradius (see, e.g, [3] and [26]).

We now prove, in the generic smooth non convex case, convergence to the set of stationary points with a rate of $O(\frac{1}{\sqrt{k}})$ for $\|\pi(T_\Omega(\tilde{x}_i), -\nabla f(\tilde{x}_i))\|$.

Theorem 4.2. *Let us consider the sequence $\{x_k\}$ generated by Algorithm 2 and assume that*

- *the angle condition* (3.1) *holds;*
- *the SSC procedure always terminates in a finite number of steps.*

Then $\{f(x_k)\}$ is decreasing, $f(x_k) \to \tilde{f} \in \mathbb{R}$ and the limit points of $\{x_k\}$ are stationary. Furthermore, for any sequence $\{\tilde{x}_k\}$ satisfying (4.6) *and $K = \tau/(L(1+\tau))$,*

$$\min_{0 \leq i \leq k} \|\pi(T_\Omega(\tilde{x}_i), -\nabla f(\tilde{x}_i))\| \leq \min_{0 \leq i \leq k} \frac{\|x_{i+1} - x_i\|}{K} \leq \sqrt{\frac{2(f(x_0) - \tilde{f})}{K^2 L(k+1)}}. \tag{4.14}$$

Similarly to what we did for Theorem 4.1, here we give a corollary for Theorem 4.2 specialized to the FW variants described in Section 3.1 (see also Table 2).

Corollary 4.2. *Let us assume that $\Omega = \mathrm{conv}(A)$, with $|A| < +\infty$ in Problem* (2.1)*. Then the sequence $\{x_k\}$ generated by Algorithm 2 with AFW (PFW or FDFW) in the SSC converges at a rate given by equation* (4.14)*, with $\tau = \tau_p/2$ (τ_p or $\tau_v/2$, respectively).*

Proof. Finite termination of the SSC follows by Proposition 4.2 and Proposition 4.3, and the angle condition is satisfied by Proposition 3.1. Thus we have all the assumptions to apply Theorem 4.2. □

Remark 4.3. *Let $G : \Omega \to \mathbb{R}_{\geq 0}$ be the FW gap (see, e.g., [38]):*

$$G(x) = \max_{s \in \Omega} \langle -\nabla f(x), s - x \rangle. \tag{4.15}$$

Then, for any $y \in \Omega$

$$G(y) = \max_{s \in \Omega} \langle -\nabla f(y), s - y \rangle = \max_{s \in \Omega \setminus \{y\}} \|s - y\| \langle -\nabla f(y), \frac{s - y}{\|s - y\|} \rangle \leq D \|\pi(T_\Omega(y), -\nabla f(y))\|, \tag{4.16}$$

where the inequality follows from Proposition 2.2.

Taking into account equation (4.16), it is easy to see that our rate is an improvement of the ones proved in [15] and [38] (see Table 2). Furthermore, we do not need to start from a vertex to avoid dependence from the support of $\{x_0\}$ like in [15, Theorem 5.1]. Finally, our method improves the conditional gradient sliding rate (NCGS) not only in LMO but also in gradients, given that from $\Omega - \{y\} \subset T_\Omega(y)$ it follows $\pi(y) \leq \|\pi(T_\Omega(y), -\nabla f(y))\|/2L$ for every $y \in \Omega$.

Algorithm	Article	LMO c.r.	Gradient c.r.	Gap
NCGS	[52]	$O\left(\frac{1}{k^{0.25}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\leq i\leq k}\pi(x_i)$
AFW, FW	[15], [38]	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\leq i\leq k}G(x_i)$
AFW, PFW, FDFW + SSC	Ours	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$	$\min_{0\leq i\leq k}\|\pi(T_\Omega(\tilde{x}_i),-\nabla f(\tilde{x}_i))\|$

Table 2: Comparison between convergence rates in the generic smooth non convex case. See also Remark 4.3. $\pi(x)=\|x-\pi\left(\Omega,x-\frac{\nabla f(x)}{2L}\right)\|$, G is the FW gap (see (4.15)).

5 Conclusions

FW variants rely on the choice of good feasible descent directions, for which there needs to be a trade-off between slope and maximal stepsize. To address this issue we proposed the SSC procedure, which allowed us to prove bad step free convergence rates under an angle condition for the directions selected by the method.

Future research directions include employing our framework to design and analyze other projection free first order methods, investigating active set identification properties of FW variants with the SSC, generalizing our framework to constrained stochastic optimization, as well as applications for the solution of real-world data science problems.

6 Appendix

We report here the missing proofs. We first state a preliminary result needed to prove Proposition 2.2:

Proposition 6.1. *Let C be a closed convex cone. For every $y\in\mathbb{R}^n$*

$$\operatorname{dist}(C^*,y)=\sup_{c\in C}\langle\hat{c},y\rangle\,.$$

As stated in [18] this is an immediate consequence of the Moreau-Yosida decomposition:

$$y=\pi(C,y)+\pi(C^*,y)\,.$$

Proposition 2.2. We first prove that

$$\sup_{h\in\Omega/\{\bar{x}\}}\left(g,\frac{h-\bar{x}}{\|h-\bar{x}\|}\right)=\sup_{h\in T_\Omega(\bar{x})\setminus\{0\}}(g,\hat{h})\,. \tag{6.1}$$

Let $h\in T_\Omega(\bar{x})\setminus\{0\}$. Then there exists sequences $\{\lambda_i\}$ and $\{h_i\}$ in $\mathbb{R}_{>0}$ and Ω respectively such that $\lambda_i(h_i-\bar{x})\to h$. In particular $\|\lambda_i(h_i-\bar{x})\|\to\|h\|$ so that we also have $\lambda_i(h_i-\bar{x})/\|\lambda_i(h_i-\bar{x})\|=(h_i-\bar{x})/\|h_i-\bar{x}\|\to\hat{h}$. Hence

$$\operatorname{cl}\left(\left\{\frac{h-\bar{x}}{\|h-\bar{x}\|}\mid h\in\Omega\setminus\{\bar{x}\}\right\}\right)=\{\hat{h}\mid h\in T_\Omega(\bar{x})/\{0\}\}\,,$$

and (6.1) follows immediately by the continuity of $(g,\cdot)$.
Since $N_\Omega(\bar{x})=T_\Omega(\bar{x})^*$ the first equality is exactly the one of Proposition 6.1 if $g\notin N_\Omega(\bar{x})$, and it is trivial since both terms are clearly 0 if $g\in N_\Omega(\bar{x})$.
It remains to prove

$$\operatorname{dist}(N_\Omega(\bar{x}),g)=\|\pi(T_\Omega(\bar{x}),g)\|\,,$$

which is true by the Moreau - Yosida decomposition. □

Proposition 4.1. Let $B_j = \bar{B}_{\langle g,\hat{d}_j\rangle/L}(x_k)$ and let T be such that $x_{k+1} = y_T$.
Inequality (4.3) applied with $j = T$ gives (4.5). Moreover, by taking $\tilde{x}_k = y_{\tilde{T}}$ for some $\tilde{T} \in [0:T]$ the conditions

$$f(x_{k+1}) \leq f(\tilde{x}_k) \leq f(x_k) - \frac{L}{2}\|x_k - \tilde{x}_k\|^2 \tag{6.2}$$

are satisfied by Lemma 4.1 and (4.3).
Let now $p_j = \|\pi(T_\Omega(y_j), -\nabla f(y_j))\|$ and $\tilde{p}_j = \|\pi(T_\Omega(y_j), g)\| = \|\pi(T_\Omega(y_j), -\nabla f(x_k))\|$. We have

$$|p_j - \tilde{p}_j| \leq L\|y_j - x_k\|\,, \tag{6.3}$$

reasoning as for (3.16). We now distinguish four cases according to how the SSC terminates.
Case 1: $T = 0$ or $d_T = 0$. Since there are no descent directions $x_{k+1} = y_T$ must be stationary for the gradient g. Equivalently, $\tilde{p}_T = \|\pi(T_\Omega(x_{k+1}), g)\| = 0$. We can now write

$$\|x_{k+1} - x_k\| \geq \frac{1}{L}(|p_T - \tilde{p}_T|) = \frac{p_T}{L} > Kp_T\,,$$

where we used (6.3) in the first inequality and $\tilde{p}_T = 0$ in the equality. Finally, it is clear that if $T = 0$ then $d_0 = 0$, since y_0 must be stationary for $-g$.
Before examining the remaining cases we remark that if the SSC terminates in Phase II then $\alpha_{T-1} = \beta_{T-1}$ must be maximal w.r.t. the conditions $y_T \in B_{T-1}$ or $y_T \in \bar{B}$. If $\alpha_{T-1} = 0$ then $y_{T-1} = y_T$, and in this case we cannot have $y_{T-1} \in \partial\bar{B}$, otherwise the SSC would terminate in Phase II of the previous cycle. Therefore necessarily $y_T = y_{T-1} \in \mathrm{int}(B_{T-1})^c$ (Case 2). If $\beta_{T-1} = \alpha_{T-1} > 0$ we must have $y_{T-1} \in \Omega_{T-1} = B_{T-1} \cap \bar{B}$, and $y_T \in \partial B_{T-1}$ (case 3) or $y_T \in \partial\bar{B}$ (case 4) respectively.
Case 2: $y_{T-1} = y_T \in \mathrm{int}(B_{T-1})^c$. We can rewrite the condition as

$$\langle g, \hat{d}_{T-1}\rangle \leq L\|y_{T-1} - x_k\| = L\|y_T - x_k\|\,. \tag{6.4}$$

Thus

$$p_T = p_{T-1} \leq \tilde{p}_{T-1} + L\|y_T - x_k\| \leq \frac{1}{\tau}\langle g, \hat{d}_{T-1}\rangle + L\|y_T - x_k\| \leq \left(\frac{L}{\tau} + L\right)\|y_T - x_k\|\,, \tag{6.5}$$

where in the equality we used $y_T = y_{T-1}$, the first inequality follows from (6.3) and again $y_T = y_{T-1}$, the second from $\frac{\langle g,\hat{d}_T\rangle}{\tilde{p}_T} \geq \mathrm{DSB}_{\mathcal{A}}(\Omega, y_T, g) \geq \mathrm{SB}_{\mathcal{A}}(\Omega) = \tau$, and the third from (6.4). Then $\tilde{x}_k = x_{k+1} = y_T$ satisfies the desired conditions.
Case 3: $y_T = y_{T-1} + \beta_{T-1}d_{T-1}$ and $y_T \in \partial B_{T-1}$. Then from $y_{T-1} \in B_{T-1}$ it follows

$$L\|y_{T-1} - x_k\| \leq \langle g, \hat{d}_{T-1}\rangle\,, \tag{6.6}$$

and $y_T \in \partial B_{T-1}$ implies

$$\langle g, \hat{d}_{T-1}\rangle = L\|y_T - x_k\|\,. \tag{6.7}$$

Combining (6.6) with (6.7) we obtain

$$L\|y_{T-1} - x_k\| \leq L\|y_T - x_k\|\,. \tag{6.8}$$

Thus

$$p_{T-1} \leq \tilde{p}_{T-1} + L\|y_{T-1} - x_k\| \leq \frac{1}{\tau}\langle g, \hat{d}_{T-1}\rangle + L\|y_{T-1} - x_k\| \leq \left(\frac{L}{\tau} + L\right)\|y_T - x_k\|\,,$$

where we used (6.7), (6.8) in the last inequality and the rest follows reasoning as for (6.5). In particular we can take $\tilde{x}_k = y_{T-1}$.
Case 4: $y_T = y_{T-1} + \beta_{T-1}d_{T-1}$ and $y_T \in \partial\bar{B}$.
The condition $x_{k+1} = y_T \in \bar{B}$ can be rewritten as

$$L\|x_{k+1} - x_k\|^2 - \langle g, x_{k+1} - x_k\rangle = 0\,. \tag{6.9}$$

For every $j \in [0:T]$ we have

$$x_{k+1} = y_j + \sum_{i=j}^{T-1} \alpha_i d_i \,. \tag{6.10}$$

We now want to prove that for every $j \in [0:T]$

$$\|x_{k+1} - x_k\| \geq \|y_j - x_k\| \,. \tag{6.11}$$

Indeed, we have

$$\begin{aligned} L\|x_{k+1} - x_k\|^2 = \langle g, x_{k+1} - x_k\rangle = \langle g, y_j - x_k\rangle + \sum_{i=j}^{T-1} \alpha_i \langle g, d_i\rangle \\ \geq \langle g, y_j - x_k\rangle \geq L\|y_j - x_k\|^2 \,, \end{aligned}$$

where we used (6.9) in the first equality, (6.10) in the second, $\langle g, d_j\rangle \geq 0$ for every j in the first inequality and $y_j \in \bar{B}$ in the second inequality.
We also have

$$\begin{aligned} &\frac{\langle g, x_{k+1} - x_k\rangle}{\|x_{k+1} - x_k\|} = \frac{\langle g, \sum_{j=0}^{T-1} \alpha_j d_j\rangle}{\|\sum_{j=0}^{T-1} \alpha_j d_j\|} \geq \frac{\langle g, \sum_{j=0}^{T-1} \alpha_j d_j\rangle}{\sum_{j=0}^{T-1} \alpha_j \|d_j\|} \\ &\geq \min\left\{ \frac{\langle g, d_j\rangle}{\|d_j\|} \mid 0 \leq j \leq T-1 \right\}. \end{aligned} \tag{6.12}$$

Thus for $\tilde{T} \in \operatorname{argmin}\left\{ \frac{\langle g, d_j\rangle}{\|d_j\|} \mid 0 \leq j \leq T-1 \right\}$

$$\langle g, \hat{d}_{\tilde{T}}\rangle \leq \frac{\langle g, x_{k+1} - x_k\rangle}{\|x_{k+1} - x_k\|} = L\|x_{k+1} - x_k\| \,, \tag{6.13}$$

where we used (6.12) in the first inequality and (6.9) in the second.
We finally have

$$p_{\tilde{T}} \leq \tilde{p}_{\tilde{T}} + L\|y_{\tilde{T}} - x_k\| \leq \frac{1}{\tau}\langle g, \hat{d}_{\tilde{T}}\rangle + L\|y_{\tilde{T}} - x_k\| \leq \left(\frac{L}{\tau} + L\right)\|x_{k+1} - x_k\| \,,$$

where we used (6.11), (6.13) in the last inequality and the rest follows reasoning as for (6.5). In particular $\tilde{x}_k = y_{\tilde{T}}$ satisfies the desired properties. □

We now prove Theorem 4.1. We start by recalling Karamata's inequality ([32], [33]) for concave functions. Given $A, B \in \mathbb{R}^N$ it is said that A majorizes B, written $A \succ B$, if

$$\begin{aligned} &\sum_{i=1}^{j} A_i \geq \sum_{i=1}^{j} B_i \text{ for } j \in [1:N] \,, \\ &\sum_{i=1}^{N} A_i = \sum_{i=1}^{N} B_i \,. \end{aligned}$$

If h is concave and $A \succ B$ by Karamata's inequality

$$\sum_{i=1}^{N} h(A_i) \leq \sum_{i=1}^{N} h(B_i) \,.$$

Theorem 4.1. If the sequence $\{x_k\}$ is finite, with $x_m = \tilde{x}$ stationary for some $m \geq 0$, we define $x_k = x_m$ for every $k \geq m$, so that we can always assume $\{x_k\}$ infinite. Notice that with this convention the sufficient decrease condition (4.5) is still satisfied for every k. Let $f_k = f(x_k) - f(x^*)$. $\{f_k\}$ is monotone

decreasing by (4.5), and nonnegative since (2.2) holds for every x_k.
We want prove $f_{k+1} \leq q f_k$. This is clear if $f_{k+1} = 0$. Otherwise using the notation of Proposition 4.1 we have

$$f_k - f_{k+1} \geq \frac{L}{2}\|x_k - x_{k+1}\|^2 \geq \frac{LK^2}{2}\|\pi(T_\Omega(\tilde{x}_k), -\nabla f(x_k))\|\,, \tag{6.14}$$

where we used (4.5) in the first inequality, (4.6) in the second. Since $\tilde{x}_k \in \{y_j\}_{j=0}^T$ by Proposition 4.1, we can apply (2.2) in $\tilde{x}_k$ to obtain

$$\frac{LK^2}{2}\|\pi(T_\Omega(\tilde{x}_k), -\nabla f(x_k))\| \geq \mu L K^2 (f(\tilde{x}_k) - f(x^*)) \geq \mu L K^2 f_{k+1}. \tag{6.15}$$

Concatenating (6.14), (6.15) and rearranging we obtain

$$f_{k+1} \leq (1 + \mu L K^2)^{-1} f_k = q f_k\,. \tag{6.16}$$

Thus by induction for any $i \geq 0$

$$f_{k+i} \leq q^i f_k\,, \tag{6.17}$$

which implies in particular (4.9).
For $i \geq 2$ let $n(i,k) = \min\{j \geq 0 \mid f_{k+j+1} < q^i f_k\}$, so that by (6.17) we have $n(i,k) \leq i$. Define $w^*_{ik}, v_{ik} \in \mathbb{R}^i_{\geq 0}$ by

$$\begin{aligned} v_{ik} &= (f_k - q f_k, ..., q^{i-1} f_k - q^i f_k)\,, \\ w^*_{ik} &= (f_k - f_{k+1}, ..., f_{n(i,k)-1} - f_{n(i,k)}, f_{n(i,k)} - q^i f_k, 0, ..., 0)\,. \end{aligned} \tag{6.18}$$

Then for $0 \leq l \leq n(i,k) - 1$ we have

$$\sum_{j=0}^{l} (v_{ik})_j = f_k - q^{l+1} f_k \leq f_k - f_{k+l+1} = \sum_{j=0}^{l} (w^*_{ik})_j\,, \tag{6.19}$$

where we used (6.17) with $l+1$ instead of i in the inequality, while for $n(i,k) \leq l \leq i-1$ we have

$$\sum_{j=0}^{l} (v_{ik})_j = f_k - q^{l+1} f_k \leq f_k - q^i f_k = \sum_{j=0}^{l} (w^*_{ik})_j\,, \tag{6.20}$$

with equality for $l = i - 1$.
Now if w_{ik} is the permutation in descreasing order of w^*_{ik} clearly $\sum_{j=0}^{l} (w^*_{ik})_j \leq \sum_{j=0}^{l} (w_{ik})_j$ for every l, and combining this with (6.19) and (6.20) we have $w_{ik} \succ v_{ik}$. Then

$$\begin{aligned} \sqrt{f_{n(i,k)} - q^i f_k} + \sum_{j=0}^{n(i,k)} \sqrt{f_{k+j} - f_{k+j+1}} &= \sum_{j=0}^{i} \sqrt{(w^*_{ik})_j} = \sum_{j=0}^{i} \sqrt{(w_{ik})_j} \leq \sum_{j=0}^{i} \sqrt{(v_{ik})_j} \\ &\leq \sqrt{f_k} \sum_{j=0}^{+\infty} \sqrt{q^j - q^{j+1}} = \frac{\sqrt{f_k(1-q)}}{1 - \sqrt{q}}\,, \end{aligned} \tag{6.21}$$

where the first inequality follows from Karamata's inequality. Taking the limit for $i \to \infty$ in the LHS we obtain

$$\sum_{j=0}^{+\infty} \sqrt{f_{k+j} - f_{k+j+1}} \leq \frac{\sqrt{f_k(1-q)}}{1 - \sqrt{q}}\,. \tag{6.22}$$

We can now bound the length of the tails of $\{x_k\}$:

$$\sum_{j=0}^{+\infty} \|x_{k+j} - x_{k+j+1}\| \leq \sqrt{\frac{2}{L}} \sum_{j=0}^{+\infty} \sqrt{f_{k+j} - f_{k+j+1}} \leq \frac{\sqrt{2 f_k (1-q)}}{\sqrt{L}(1 - \sqrt{q})} \leq \frac{\sqrt{2 f_0 (1-q)}}{\sqrt{L}(1 - \sqrt{q})} q^{\frac{k}{2}}\,, \tag{6.23}$$

where we used (4.5) in the first inequality, (6.22) in the second and (6.17) in the third. In particular $x_k \to \tilde{x}^*$ with

$$\|x_k - \tilde{x}^*\| \leq \sum_{j=0}^{+\infty} \|x_{k+j} - x_{k+j+1}\| = \frac{\sqrt{2f_0(1-q)}}{\sqrt{L}(1-\sqrt{q})} q^{\frac{k}{2}} \tag{6.24}$$

by (6.23). □

Theorem 4.2. The sequence $\{f(x_k)\}$ is decreasing by (4.5). Thus by compactness $f(x_k) \to \tilde{f} \in \mathbb{R}$ and in particular $f(x_k) - f(x_{k+1}) \to 0$. So that by (4.5) also $\|x_{k+1} - x_k\| \to 0$. Let $\{x_{k(i)}\} \to \tilde{x}^*$ be any convergent subsequence of $\{x_k\}$. For $\{\tilde{x}_k\}$ chosen as in the proof of Proposition 4.1 we have $\|\tilde{x}_k - x_k\| \leq \|x_{k+1} - x_k\|$ because $\tilde{x}_k = y_T = x_k$ in case 1 and case 2, by (6.8) in case 3, and by (6.11) in case 4. Therefore

$$\|\tilde{x}_{k(i)} - x_{k(i)}\| \leq \|x_{k(i)+1} - x_{k(i)}\| \to 0\,.$$

Furthermore, $\|\pi(T_\Omega(\tilde{x}_{k(i)}), -\nabla f(\tilde{x}_{k(i)}))\| \leq \frac{\|x_{k(i)+1} - x_{k(i)}\|}{K} \to 0$ again by Proposition 4.1, so that $\tilde{x}_{k(i)} \to \tilde{x}^*$ with $\|\pi(T_\Omega(\tilde{x}_{k(i)}), -\nabla f(\tilde{x}_{k(i)}))\| \to 0$. Then $\|\pi(T_\Omega(\tilde{x}^*), -\nabla f(\tilde{x}^*))\| = 0$ and $\tilde{x}^*$ is stationary.

The first inequality in (4.14) follows directly from (4.6). As for the second, we have

$$\begin{aligned}
&\frac{k+1}{K^2}(\min_{0\leq i\leq k} \|x_{i+1} - x_i\|)^2 = \frac{k+1}{K^2} \min_{0\leq i\leq k} \|x_{i+1} - x_i\|^2 \\
\leq &\frac{1}{K^2} \sum_{i=0}^{k} \|x_i - x_{i+1}\|^2 \leq \frac{2}{LK^2} \sum_{i=0}^{k} (f(x_{i+1}) - f(x_i)) \leq \frac{2(f(x_0) - \tilde{f})}{LK^2}\,,
\end{aligned}$$

where we used (4.5) in the first inequality, $\{f(x_i)\}$ decreasing together with $f(x_i) \to \tilde{f}$ in the second and the thesis follows by rearranging terms. □

References

[1] P-A Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[2] P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[3] Ralph Alexander. The width and diameter of a simplex. *Geometriae Dedicata*, 6(1):87–94, 1977.

[4] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[5] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[6] M. V. Balashov, B. T. Polyak, and A. A. Tremba. Gradient projection and conditional gradient methods for constrained nonconvex minimization. *Numerical Functional Analysis and Optimization*, 41(7):822–849, 2020.

[7] Mohammad Ali Bashiri and Xinhua Zhang. Decomposition-invariant conditional gradient for general polytopes with line search. In *Advances in Neural Information Processing Systems*, pages 2690–2700, 2017.

[8] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

[9] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Deep Frank-Wolfe for neural network optimization. In *International Conference on Learning Representations*, 2018.

[10] Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[11] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[12] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[13] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[14] Immanuel M Bomze, Francesco Rinaldi, and Samuel Rota Bulo. First-order methods for the impatient: Support identification in finite time with convergent Frank-Wolfe variants. *SIAM Journal on Optimization*, 29(3):2211–2226, 2019.

[15] Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Active set complexity of the away-step Frank–Wolfe algorithm. *SIAM Journal on Optimization*, 30(3):2470–2500, 2020.

[16] Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditonal gradients. In *International Conference on Machine Learning*, pages 735–743. PMLR, 2019.

[17] Gábor Braun, Sebastian Pokutta, and Daniel Zink. Lazifying conditional gradient algorithms. In *ICML*, pages 566–575, 2017.

[18] James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[19] Michael D Canon and Clifton D Cullum. A tight upper bound on the rate of convergence of Frank-Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968.

[20] Cyrille W Combettes and Sebastian Pokutta. Boosting Frank-Wolfe by chasing gradients. *arXiv preprint arXiv:2003.06369*, 2020.

[21] Andrea Cristofari, Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. An active-set algorithmic framework for non-convex optimization problems over the simplex. *Computational Optimization and Applications*, 77:57–89, 2020.

[22] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[23] Robert M Freund, Paul Grigas, and Rahul Mazumder. An extended Frank-Wolfe method with in-face directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017.

[24] Dan Garber. Revisiting Frank-Wolfe for polytopes: Strict complementary and sparsity. *arXiv preprint arXiv:2006.00558*, 2020.

[25] Luigi Grippo, Francesco Lampariello, and Stephano Lucidi. A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[26] Peter Gritzmann and Marek Lassak. Estimates for the minimal width of polytopes inscribed in convex bodies. *Discrete & Computational Geometry*, 4(6):627–635, 1989.

[27] Jacques Guelat and Patrice Marcotte. Some comments on Wolfe's away step. *Mathematical Programming*, 35(1):110–119, 1986.

[28] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[29] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, pages 427–435, 2013.

[30] Carl Johnell and Morteza Haghir Chehreghani. Frank-Wolfe optimization for dominant set clustering. *arXiv preprint arXiv:2007.11652*, 2020.

[31] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[32] Zoran Kadelburg, Dusan Dukic, Milivoje Lukic, and Ivan Matic. Inequalities of Karamata, Schur and Muirhead, and some applications. *The Teaching of Mathematics*, 8(1):31–45, 2005.

[33] Jovan Karamata. Sur une inégalité relative aux fonctions convexes. *Publications de l'Institut Mathématique*, 1(1):145–147, 1932.

[34] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[35] Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting Frank-Wolfe. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1275–1283. PMLR, 2019.

[36] Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Stationarity results for generating set search for linearly constrained optimization. *SIAM Journal on Optimization*, 17(4):943–968, 2007.

[37] Vladimir Kolmogorov. Practical Frank-Wolfe algorithms. *arXiv preprint arXiv:2010.09567*, 2020.

[38] Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[39] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28:496–504, 2015.

[40] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[41] Kfir Levy and Andreas Krause. Projection free online learning over smooth sets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1458–1466, 2019.

[42] Robert Michael Lewis, Anne Shepherd, and Virginia Torczon. Implementing generating set search methods for linearly constrained minimization. *SIAM Journal on Scientific Computing*, 29(6):2507–2530, 2007.

[43] Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

[44] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[45] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[46] Hassan Mortagy, Swati Gupta, and Sebastian Pokutta. Walking in the shadow: A new perspective on descent directions for constrained minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[47] Julie Nutini, Mark Schmidt, and Warren Hare. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13(4):645–655, 2019.

[48] Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, and Simon Lacoste-Julien. Minding the gaps for block frank-wolfe optimization of structured svms. In *International Conference on Machine Learning*, pages 593–602. PMLR, 2016.

[49] Javier Peña and Daniel Rodriguez. Polytope conditioning and linear convergence of the Frank-Wolfe algorithm. *Math. Oper. Res.*, 44(1):1–18, 2018.

[50] Fabian Pedregosa, Geoffrey Negiar, Armin Askari, and Martin Jaggi. Linearly convergent Frank-Wolfe with backtracking line-search. In *International Conference on Artificial Intelligence and Statistics*, pages 1–10. PMLR, 2020.

[51] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[52] Chao Qu, Yan Li, and Huan Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018.

[53] Luis Rademacher and Chang Shu. The smoothed complexity of Frank-Wolfe methods via conditioning of random matrices and polytopes. *arXiv preprint arXiv:2009.12685*, 2020.

[54] Francesco Rinaldi and Damiano Zeffiro. A unifying framework for the analysis of projection-free first-order methods under a sufficient slope condition. *arXiv preprint arXiv:2008.09781*, 2020.

[55] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[56] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[57] Philip Wolfe. Convergence theory in nonlinear programming. *Integer and nonlinear programming*, pages 1–36, 1970.

[58] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[59] Li Zhang, Weijun Zhou, and Dong-Hui Li. A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence. *IMA Journal of Numerical Analysis*, 26(4):629–640, 2006.