Active Set Complexity of the Away-Step Frank--Wolfe Algorithm

# ACTIVE SET COMPLEXITY OF THE AWAY–STEP FRANK–WOLFE ALGORITHM

IMMANUEL M. BOMZE*, FRANCESCO RINALDI†, AND DAMIANO ZEFFIRO‡

**Abstract.** In this paper, we study active set identification results for the away-step Frank-Wolfe algorithm in different settings. We first prove a local identification property that we apply, in combination with a convergence hypothesis, to get an active set identification result. We then prove, in the nonconvex case, a novel $O(1/\sqrt{k})$ convergence rate result and active set identification for different stepsizes (under suitable assumptions on the set of stationary points). By exploiting those results, we also give explicit active set complexity bounds for both strongly convex and nonconvex objectives. While we initially consider the probability simplex as feasible set, in the appendix we show how to adapt some of our results to generic polytopes.

**Key words.** Surface Identification, Manifold Identification, Active Set Complexity

**AMS subject classifications.** 65K05, 90C06, 90C30

**1. Introduction.** Identifying a surface containing a solution (and/or the support of sparse solutions) represents a relevant task in optimization, since it allows to reduce the dimension of the problem at handle and to apply a more sophisticated method in the end (see, e.g. [4, 6, 14, 15, 16, 20, 21, 22]). This is the reason why, in the last decades, identification properties of optimization methods have been the subject of extensive studies.

The Frank-Wolfe (FW) algorithm, first introduced in [17], is a classic first-order optimization method that has recently re-gained popularity thanks to the way it can easily handle the structured constraints appearing in many real-world applications. This method and its variants have been indeed applied in the context of, e.g., submodular optimization problems [1], variational inference problems [26] and sparse neural network training [18]. It is important to notice that the FW approach has a relevant drawback with respect to other algorithms: even when dealing with the simplest polytopes, it cannot identify the active set in finite time (see, e.g., [8]). Due to the renewed interest in the method, it has hence become a relevant issue to determine whether some FW variants admit active set identification properties similar to those of other first order methods. In this paper we focus on the away-step Frank-Wolfe (AFW) method and analyze active set identification properties for problems of the form

$$\min \{f(x) \mid x \in \Delta_{n-1}\},$$

where the objective $f$ is a differentiable function with Lipschitz regular gradient and the feasible set

$$\Delta_{n-1} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1,\, x \geq 0 \right\}$$

is the probability simplex. We further extend some of the active set complexity results to general polytopes.

**1.1. Contributions.** It is a classic result that on polytopes and under strict complementarity conditions the AFW with exact linesearch identifies the face containing the minimum in finite time for strongly convex objectives [19]. More general active set identification properties for Frank-Wolfe variants have recently been analyzed in [8], where the authors proved active set identification for sequences convergent to a stationary point, and AFW convergence to a stationary point for $C^2$ objectives with a finite number of stationary points and satisfying a technical convexity-concavity assumption (this assumption is substantially a generalization of a property related to quadratic possibly indefinite functions). The main contributions of this article with respect to [8] are twofold:

*ISOR, VCOR & ds:UniVie, Universität Wien, Austria (immanuel.bomze@univie.ac.at)
†Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (rinaldi@math.unipd.it)
‡Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (damiano.zeffiro@math.unipd.it)

- First, we give quantitative local and global active set identification complexity bounds under suitable assumptions on the objective. The key element in the computation of those bounds is a quantity that we call "active set radius". This radius determines a neighborhood of a stationary point for which the AFW at each iteration identifies an active constraint (if there is any not yet identified one). In particular, to get the active set complexity bound it is sufficient to know how many iterations it takes for the AFW sequence to enter this neighborhood.

- Second, we analyze the identification properties of AFW without the technical convexity-concavity $C^2$ assumption used in [8] (we consider general nonconvex objectives with Lipschitz gradient instead). More specifically, we prove active set identification under different conditions on the stepsize and some additional hypotheses on the support of stationary points.

In order to prove our results, we consider stepsizes dependent on the Lipschitz constant of the gradient (see, e.g., [2], [24] and references therein). By exploiting the affine invariance property of the AFW (see, e.g., [25]), we also extend some of the results to generic polytopes. In our analysis we will see how the AFW identification properties are related to the value of Lagrangian multipliers on stationary points. This, to the best of our knowledge, is the first time that some active set complexity bounds are given for a variant of the FW algorithm.

The paper is organized as follows: after presenting the AFW method and the setting in Section 2, we study the local behaviour of this algorithm regarding the active set in Section 3. In Section 4 we provide active set identification results in a quite general context, and apply these to the strongly convex case for obtaining complexity bounds. Section 5 treats the nonconvex case, giving both global and local active set complexity bounds. In the final Section 6 we draw some conclusions. To improve readability, some technical details are deferred to an appendix.

**1.2. Related work.** In [9] the authors proved that the projected gradient method and other converging sequential quadratic programming methods identify quasi-polyhedral faces under some non-degeneracy conditions. In [10] those results were extended to the case of exposed faces in polyhedral sets without the nondegeneracy assumptions. This extension is particularly relevant to our work since the identification of exposed faces in polyhedral sets is the framework that we will use in studying the AFW on polytopes. In [35] the results of [9] were generalized to certain nonpolyhedral surfaces called "$C^p$ identifiable" contained in the boundary of convex sets. A key insight in these early works was the openness of a generalized normal cone defined for the identifiable surface containing a nondegenerate stationary point. This openness guarantees that, in a neighborhood of the stationary point, the projection of the gradient identifies the related surface. It turns out that for linearly constrained sets the generalized normal cone is related to positive Lagrangian multipliers on the stationary point.
A generalization of [9] to nonconvex sets was proved in [11], while an extension to nonsmooth objectives was first proved in [23]. Active set identification results have also been proved for a variety of projected gradient, proximal gradient and stochastic gradient related methods (see for instance [33] and references therein).
Recently, explicit active set complexity bounds have been given for some of the methods listed above. Bounds for proximal gradient and block coordinate descent method were analyzed in [31] and [30] under strong convexity assumptions on the objective. A more systematic analysis covering many gradient related proximal methods (like, e.g., accelerated gradient, quasi Newton and stochastic gradient proximal methods) was carried out in [33].
As for FW-like methods, in addition to the results in [19] and [8] discussed earlier, identification results have been proved in [13] for fully corrective variants on the probability simplex. However, since fully corrective variants require to compute the minimum of the objective on a given face at each iteration, they are not suited for nonconvex problems.

**2. Preliminaries.** In the rest of this article $f : \Delta_{n-1} \to \mathbb{R}$ will be a function with gradient having Lipschitz constant $L$ and $\mathcal{X}^*$ will be the set of stationary points of $f$. The constant $L$ will also be used as Lipschitz constant for $\nabla f$ with respect to the norm $\|\cdot\|_1$. This does not require any additional

hypothesis on $f$ since $\|\cdot\|_1 \geq \|\cdot\|$, so that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \leq L\|x-y\|_1$$

for every $x, y \in \Delta_{n-1}$.
For $x \in \mathbb{R}^n$, $X \subset \mathbb{R}^n$ the function $\operatorname{dist}(x, X)$ will be the standard point set distance and for $A \subset \mathbb{R}^n$ the function $\operatorname{dist}(A, X)$ will be the minimal distance between points in the sets:

$$\operatorname{dist}(A, X) = \inf_{a \in A, x \in X} \|a - x\| .$$

We define $\operatorname{dist}_1$ in the same way but with respect to $\|\cdot\|_1$. We use the notation

$$\operatorname{supp}(x) = \{i \in [1:n] \mid x_i \neq 0\}$$

for the support of a point $x \in \mathbb{R}^n$.
Given a (convex and bounded) polytope $P$ and a vector $c$ we define the face of $P$ *exposed by* $c$ as

$$\mathcal{F}(c) = \operatorname{argmax}\{c^\top x \mid x \in P\} .$$

It follows from the definition that the face of $P$ exposed by a linear function is always unique and nonempty.
We now introduce the multiplier functions, which were recently used in [14] to define an active set strategy for minimization over the probability simplex.
For every $x \in \Delta_{n-1}$, $i \in [1:n]$ the multiplier function $\lambda_i : \Delta_{n-1} \to \mathbb{R}$ is defined as

$$\lambda_i(x) = \nabla f(x)^\top (e_i - x),$$

or in vector form

$$\lambda(x) = \nabla f(x) - x^\top \nabla f(x) e .$$

For every $x \in \mathcal{X}^*$ these functions coincide with the Lagrangian multipliers of the constraints $x_i \geq 0$.
For a sequence $\{a_k\}_{k \in \mathbb{N}_0}$ we will drop the subscript and write simply $\{a_k\}$ (unless of course the sequence is defined on some other index set).
FW variants require a linear minimization oracle for the feasible set (the probability simplex in our case):

$$\operatorname{LMO}_{\Delta_{n-1}}(r) \in \operatorname{argmin}\{x^\top r \mid x \in \Delta_{n-1}\}.$$

Keeping in mind that

$$\Delta_{n-1} = \operatorname{conv}(\{e_i,\ i = 1, \ldots, n\}),$$

we can assume that $\operatorname{LMO}_{\Delta_{n-1}}(r)$ always returns a vertex of the probability simplex, that is

$$\operatorname{LMO}_{\Delta_{n-1}}(r) = e_{\hat{\imath}}$$

with $\hat{\imath} \in \operatorname{argmin}_i r_i$.

Algorithm 1 is the classical FW method on the probability simplex. At each iteration, this first order method generates a descent direction that points from the current iterate $x_k$ to a vertex $s_k$ minimizing the scalar product with the gradient, and then moves along this search direction of a suitable stepsize if stationarity conditions are not satisfied.

| **Algorithm 1** Frank–Wolfe method on the probability simplex |
|---|
| 1. **Initialize** $x_0 \in \Delta_{n-1}$, $k := 0$ |
| 2. Set $s_k := e_{\hat{\imath}}$, with $\hat{\imath} \in \operatorname{argmin}_i \nabla_i f(x_k)$ and $d_k^{\mathcal{FW}} := s_k - x_k$ |
| 3. If $x_k$ is stationary, then STOP |
| 4. Choose the step size $\alpha_k \in (0, 1]$ with a suitable criterion |
| 5. Update: $x_{k+1} := x_k + \alpha_k d_k^{\mathcal{FW}}$ |
| 6. Set $k := k + 1$. Go to Step 2. |

It is well known [12, 34] that the method exhibits a zig zagging behaviour as the sequence of iterates $\{x_k\}$ approaches a solution on the boundary of the feasible set. In particular, when this happens the sequence $\{x_k\}$ converges slowly and, as we already mentioned, it does not identify the smallest face containing the solution in finite time. Both of these issues are solved by the away-step variant of the FW method, reported in Algorithm 2. The AFW at every iteration chooses between the classic FW direction and the away-step direction $d_k^{\mathcal{A}}$ calculated at Step 4. This away direction shifts weight away from the worst vertex to the other vertices used to represent the iterate $x_k$. Here the worst vertex (among those having positive weight in the iterate representation) is the one with the greatest scalar product with the gradient, or, equivalently, the one that maximizes the linear approximation of $f$ given by $\nabla f(x_k)$. The stepsize upper bound $\alpha_k^{\max}$ in Step 8 is the maximal possible for the away direction given the boundary conditions. When the algorithm performs an away step, we have that either the support of the current iterate stays the same or decreases of one (we get rid of the component whose index is associated to the away direction in case $\alpha_k = \alpha_k^{\max}$). On the other hand, when the algorithm performs a Frank Wolfe step, only the vertex given by the LMO is eventually added to the support of the current iterate. These two properties are fundamental for the active set identification of the AFW.

**Algorithm 2** Away–step Frank–Wolfe on the probability simplex

1. **Initialize** $x_0 \in \Delta_{n-1}$, $k := 0$
2. Set $s_k := e_{\hat{\imath}}$, with $\hat{\imath} \in \operatorname{argmin}_i \nabla_i f(x_k)$ and $d_k^{\mathcal{FW}} := s_k - x_k$
3. If $x_k$ is stationary then STOP
4. Let $v_k := e_{\hat{\jmath}}$, with $\hat{\jmath} \in \operatorname{argmax}_{j \in S_k} \nabla_j f(x_k)$, $S_k := \{j : (x_k)_j > 0\}$ and $d_k^{\mathcal{A}} := x_k - v_k$
5. If $-\nabla f(x_k)^\top d_k^{\mathcal{FW}} \geq -\nabla f(x_k)^\top d_k^{\mathcal{A}}$ then
6. $\quad d_k := d_k^{\mathcal{FW}}$, and $\alpha_k^{\max} := 1$
7. else
8. $\quad d_k := d_k^{\mathcal{A}}$, and $\alpha_k^{\max} := (x_k)_i/(1-(x_k)_i)$
9. End if
10. Choose the step size $\alpha_k \in (0, \alpha_k^{\max}]$ with a suitable criterion
11. Update: $x_{k+1} := x_k + \alpha_k d_k$
12. $k := k+1$. Go to step 2.

In our analysis, we will sometimes require a lower bound on the step size which is always satisfied by the exact linesearch and the Armijo rule for a proper choice of the parameters.

**3. Local active set variables identification property of the AFW.** In this section we prove a rather technical proposition which is the key tool to give quantitative estimates for the active set complexity. It states that when the sequence is close enough to a fixed stationary point at every step the AFW identifies one variable violating the complementarity conditions with respect to the multiplier functions on this stationary point (if it exists), and it sets the variable to 0 with an away step. The main difficulty is giving a tight estimate for how close the sequence must be to a stationary point for this identifying away step to take place.

A lower bound on the size of the nonmaximal away steps is needed in the following theorem, otherwise of course the steps could be arbitrarily small and there could be no convergence at all.

Let $\{x_k\}$ be the sequence of points generated by the AFW. We further indicate with $x^*$ a fixed point in $\mathcal{X}^*$, with *the extended support*

$$I = \{i \in [1:n] \mid \lambda_i(x^*) = 0\}$$

and with $I^c = \{1, \ldots n\} \setminus I$. Note that by complementary slackness, we have $x_j^* = 0$ for all $j \in I^c$.

Before proving the main theorem we need to prove the following lemma to bound the Lipschitz constant of the multipliers on stationary points.

LEMMA 3.1. *Given $h > 0$, $x_k \in \Delta_{n-1}$ such that $\|x_k - x^*\|_1 \leq h$ let*

$$O_k = \{i \in I^c \mid (x_k)_i = 0\}$$

*and assume that* $O_k \neq I^c$. *Let* $\delta_k = \max_{i\in[1:n]\setminus O_k} \lambda_i(x^*)$. *For every* $i \in \{1, ..., n\}$*:*

$$|\lambda_i(x^*) - \lambda_i(x_k)| \leq h(L + \frac{\delta_k}{2}) \ . \tag{3.1}$$

*Proof.* By considering the definition of $\lambda(x)$, we can write

$$\begin{aligned} |\lambda_i(x_k) - \lambda_i(x^*)| &= |\nabla f(x_k)_i - \nabla f(x^*)_i + \nabla f(x^*)^\top (x^* - x_k) + (\nabla f(x^*) - \nabla f(x_k))^\top x_k| \\ &\leq |\nabla f(x^*)_i - \nabla f(x_k)_i + (\nabla f(x_k) - \nabla f(x^*))^\top x_k| + |\nabla f(x^*)^\top (x^* - x_k)| \ . \end{aligned} \tag{3.2}$$

By taking into account the fact that $x_k \in \Delta_{n-1}$ and gradient of $f$ is Lipschitz continuous, we have

$$\begin{aligned} |\nabla f(x_k)_i - \nabla f(x^*)_i + (\nabla f(x^*) - \nabla f(x_k))^\top x_k| &= |(\nabla f(x^*) - \nabla f(x_k))^\top (x_k - e_i)| \\ &\leq \|\nabla f(x^*) - \nabla f(x_k)\|_1 \|x_k - e_i\|_\infty \\ &\leq Lh, \end{aligned} \tag{3.3}$$

where the last inequality is justified by the Hölder inequality with exponents $1, \infty$.
We now bound the second term in the right-hand side of (3.2). Let

$$u_j = \max\{0, (x^* - x_k)_j\}, \ l_j = \max\{0, -(x^* - x_k)_j\} \ .$$

We have $\sum_{j\in[1:n]} x^*_j = \sum_{j\in[1:n]} (x_k)_j = 1$ since $\{x^*, x_k\} \subset\in \Delta_{n-1}$, so that

$$\sum_{j\in[1:n]} (x^* - x_k)_j = \sum_{j\in[1:n]} (u_j - l_j) = 0 \quad \text{and hence} \quad \sum_{j\in[1:n]} u_j = \sum_{i\in[1:n]} l_j.$$

Moreover, $h' \stackrel{\text{def}}{=} 2\sum_{j\in[1:n]} u_j = 2\sum_{j\in[1:n]} l_j = \sum_{j\in[1:n]} u_j + l_j = \sum_{j\in[1:n]} |x^*_j - (x_k)_j| \leq h$, hence

$$h'/2 = \sum_{j\in[1:n]} u_j = \sum_{j\in[1:n]} l_j \leq h/2 \ .$$

We can finally bound the second piece of (3.2), using $u_j = l_j = 0$ for all $j \in O_k$ (because $(x_k)_j = x^*_j = 0$):

$$\begin{aligned} |\nabla f(x^*)^\top (x^* - x_k)| = |\nabla f(x^*)^\top k - \nabla f(x^*)^\top l| &\leq \frac{h'}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m) \\ &\leq \frac{h}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m), \end{aligned} \tag{3.4}$$

where $\nabla f(x_k)_M$ and $\nabla f(x_k)_m$ are respectively the maximum and minimum component of the gradient in $[1 : n] \setminus O_k$.
Now, considering inequalities (3.2), (3.3) and (3.4), we can write

$$|\lambda_i(x_k) - \lambda_i(x^*)| \leq Lh + \frac{h}{2}(\nabla f(x^*)_M - \nabla f(x^*)_m).$$

By taking into account the definition of $\delta_k$ and the fact that $\lambda(x^*)_j \geq 0$ for all $j$, we can write

$$\delta_k = \max_{i,j\in[1:n]\setminus O_k} (\nabla f(x^*)_i - \nabla f(x^*)_j) \geq \nabla f(x^*)_M - \nabla f(x^*)_m.$$

We can finally write

$$|\lambda_i(x_k) - \lambda_i(x^*)| \leq h(L + \frac{\delta_k}{2}),$$

thus concluding the proof. ∎

We now show a few simple but important results that connect the multipliers and the directions selected by the AFW algorithm. Notice that for a fixed $x_k$ the multipliers $\lambda_i(x_k)$ are the values of the linear function $x \mapsto \nabla f(x_k)^\top x$ on the vertices of $\Delta_{n-1}$ (up to a constant), which in turn are the values considered in the AFW to select the direction. This basic observation is essentially everything we need for the next results.

LEMMA 3.2. *Let* $S_k = \{i \in \{1, ..., n\} \mid (x_k)_i > 0\}$. *Then*

*(a) If* $\max\{\lambda_i(x_k) \mid i \in S_k\} > \max\{-\lambda_i(x_k) \mid i \in [1:n]\}$, *then the AFW performs an away step with* $d_k = d_k^{\mathcal{A}} = x_k - e_{\hat{\imath}}$ *for some* $i \in \operatorname{argmax}\{\lambda_i(x_k) \mid i \in S_k\}$.

*(b) For every* $i \in [1:n] \setminus S_k$ *if* $\lambda_i(x_k) > 0$ *then* $(x_{k+1})_i = (x_k)_i = 0$.

*Proof.* (a) Notice that since the vertices of the probability simplex are linearly independent for every $k$ the set of active atoms is necessarily $S_k$. In particular
$d_k^{\mathcal{A}} \in \operatorname{argmax}\{-\nabla f(x_k)^\top d) \mid d = x_k - e_i, i \in S_k\}$ and this implies

$$d_k^{\mathcal{A}} = x_k - e_{\hat{\imath}} \quad \text{for some } \hat{\imath} \in \operatorname{argmax}\{-\nabla f(x_k)^\top (x_k - e_i) \mid i \in S_k\} = \operatorname{argmax}\{\lambda_i(x_k) \mid i \in S_k\} \ . \tag{3.5}$$

As a consequence of (3.5)

$$-\nabla f(x_k)^\top d_k^{\mathcal{A}} = \max\{-\nabla f(x_k)^\top d \mid d = x_k - e_i, i \in S_k\} = \max\{\lambda_i(x_k) \mid i \in S_k\} \ , \tag{3.6}$$

where the second equality follows from $\lambda_i(x_k) = -\nabla f(x_k)^\top d$ with $d = x_k - e_i$.
Analogously

$$\begin{aligned} -\nabla f(x_k)^\top d_k^{\mathcal{FW}} &= \max\{-\nabla f(x_k)^\top d \mid d = e_i - x_k, i \in \{1, ...n\}\} = \\ &= \max\{-\lambda_i(x_k) \mid i \in \{1, ...n\}\} \ . \end{aligned} \tag{3.7}$$

We can now prove that $-\nabla f(x_k)^\top d_k^{\mathcal{FW}} < -\nabla f(x_k)^\top d_k^{\mathcal{A}}$, so that the away direction is selected under assumption (a):

$$\begin{aligned} &-\nabla f(x_k)^\top d_k^{\mathcal{FW}} = \max\{-\lambda_i(x_k) \mid i \in \{1, ...n\}\} < \\ &< \max\{\lambda_i(x_k) \mid i \in S_k\} = -\nabla f(x_k)^\top d_k^{\mathcal{A}}, \end{aligned}$$

where we used (3.6) and (3.7) for the first and the second equality respectively, and the inequality is true by hypothesis.
(b) By considering the fact that $(x_k)_i = 0$, we surely cannot choose the vertex $e_i$ to define the away-step direction. Furthermore, since $\lambda(x_k)_i = \nabla f(x_k)^\top (e_i - x_k) > 0$, direction $d = e_i - x_k$ cannot be chosen as the Frank-Wolfe direction at step $k$ as well. This guarantees that $(x_{k+1})_i = 0$. ∎

We can now prove the main theorem. The strategy will be to split $[1:n]$ in three subsets $I$, $J_k \subset I^c$ and $O_k = I^c \setminus J_k$ and use Lemma 3.1 to control the variation of the multiplier functions on each of these three subsets. In the proof we examine two possible cases under the assumption of being close enough to a stationary point. If $J_k = \emptyset$, which means that the current iteration of the AFW has identified the support of the stationary point, then we will show that the AFW chooses a direction contained in the support, so that also $J_{k+1} = \emptyset$.
If $J_k \neq \emptyset$, we will show that in the neighborhood claimed by the theorem the largest multiplier in absolute value is always positive, with index in $J_k$, and big enough, so that the corresponding away step is maximal. This means that the AFW at the iteration $k+1$ identifies a new active variable.

THEOREM 3.1. *If* $I^c$ *is not the empty set, let us define*

$$\delta_{\min} = \min\{\lambda_i(x^*) \mid i \in I^c\} > 0, \ J_k = \{i \in I^c \mid (x_k)_i > 0\} \ .$$

*Assume that for every* $k$ *such that* $d_k = d_k^{\mathcal{A}}$ *the step size* $\alpha_k$ *is either maximal with respect to the boundary condition (that is* $\alpha_k = \alpha_k^{\max}$*) or* $\alpha_k \geq \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}$. *If* $\|x_k - x^*\|_1 < \frac{\delta_{\min}}{\delta_{\min}+2L} = r_*$ *then*

$$|J_{k+1}| \leq \max\{0, |J_k| - 1\} \ . \tag{3.8}$$

*The latter relation also holds in case* $I^c = \emptyset$ *whence we put* $r_* = +\infty$.

*Proof.* If $I^c = \emptyset$, or equivalently, if $\lambda(x^*) = 0$, then there is nothing to prove since $J_k \subset I^c = \emptyset \Rightarrow |J_k| = |J_{k+1}| = 0$.
So assume $I^c \neq \emptyset$. By optimality conditions $\lambda_i(x^*) \geq 0$ for every $i$, so necessarily $\delta_{\min} > 0$.
For every $i \in [1:n]$, by Lemma 3.1

$$\begin{aligned} \lambda_i(x_k) &\geq \lambda_i(x^*) - \|x_k - x^*\|_1(L + \frac{\delta_k}{2}) > \\ &> \lambda_i(x^*) - r_*(L + \frac{\delta_k}{2}) = \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \ . \end{aligned} \tag{3.9}$$

We now distinguish two cases.
**Case 1:** $|J_k| = 0$. Then $\delta_k = 0$ because $J_k \cup I = I$ and $\lambda_i(x^*) = 0$ for every $i \in I$. Relation (3.9) becomes

$$\lambda_i(x_k) \geq \lambda_i(x^*) - \frac{\delta_{\min} L}{2L + \delta_{\min}},$$

so that for every $i \in I^c$, since $\lambda_i(x^*) \geq \delta_{\min}$, we have

$$\lambda_i(x_k) \geq \delta_{\min} - \frac{\delta_{\min} L}{2L + \delta_{\min}} > 0 \ . \tag{3.10}$$

This means that for every $i \in I^c$ we have $(x_k)_i = 0$ by the Case 1 condition $J_k = \emptyset$ and $\lambda_i(x_k) > 0$ by (3.10). We can then apply part (b) of Lemma 3.2 and conclude $(x_{k+1})_i = 0$ for every $i \in I^c$. Hence $J_{k+1} = \emptyset = J_k$ and Theorem 3.1 is proved in this case.
**Case 2.** $|J_k| > 0$. For every $i \in \operatorname{argmax}\{\lambda_j(x^*) \mid j \in J_k\}$, we have

$$\lambda_i(x^*) = \max_{j \in J_k} \lambda_j(x^*) = \max_{j \in J_k \cup I} \lambda_j(x^*),$$

where we used the fact that $\lambda_j(x^*) = 0 < \lambda_i(x^*)$ for every $j \in I$. Then by the definition of $\delta_k$, it follows

$$\lambda_i(x^*) = \delta_k.$$

Thus (3.9) implies

$$\lambda_i(x_k) > \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} = \delta_k - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}}, \tag{3.11}$$

where we used (3.9) in the inequality. But since $\delta_k \geq \delta_{\min}$ and the function $y \mapsto -\frac{y}{2L+y}$ is decreasing in $\mathbb{R}_{>0}$ we have

$$\delta_k - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \geq \delta_k - \frac{\delta_k(L + \frac{\delta_k}{2})}{2L + \delta_k} = \frac{\delta_k}{2} \ . \tag{3.12}$$

Concatenating (3.11) with (3.12), we finally obtain

$$\lambda_i(x_k) > \frac{\delta_k}{2} \ . \tag{3.13}$$

We will now show that $d_k = x_k - e_{\hat{\jmath}}$ with $\hat{\jmath} \in J_k$.
For every $j \in I$, since $\lambda_j(x^*) = 0$, again by Lemma 3.1, we have

$$\begin{aligned} |\lambda_j(x_k)| = |\lambda_j(x_k) - \lambda_j(x^*)| &\leq \|x_k - x^*\|_1(L + \delta_k/2) < \\ &< r_*(L + \delta_k/2) = \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} \leq \delta_k/2, \end{aligned} \tag{3.14}$$

where we used $\|x_k - x^*\|_1 < r_*$, which is true by definition, in the first inequality, and rearranged (3.12) to get the last inequality. For every $j \in I^c$, by (3.9), we can write

$$\lambda_j(x_k) > \delta_{\min} - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} > -\frac{\delta_k}{2} \ .$$

Then using this together with (3.14) and (3.11), we get $-\lambda_j(x_k) < \delta_k/2 < \lambda_h(x_k)$ for every $j \in [1:n], h \in \operatorname{argmax}\{\lambda_q(x^*) \mid q \in J_k\}$. So the hypothesis of Lemma 3.2 is satisfied and $d_k = d_k^{\mathcal{A}} = x_k - e_{\hat\jmath}$ with $\hat\jmath \in \operatorname{argmax}\{\lambda_j(x_k) \mid j \in S_k\}$. We need to show $\hat\jmath \in J_k$. But $S_k \subseteq I \cup J_k$ and by (3.14) if $\hat\jmath \in I$ then $\lambda_l(x_k) < \delta_k/2 < \lambda_j(x_k)$ for every $j \in \operatorname{argmax}\{\lambda_j(x^*) \mid j \in J_k\}$. If $\hat\jmath \in O_k$ then $(x_k)_{\hat\jmath} = 0$ and $\hat\jmath \notin S_k$. Hence we can conclude $\operatorname{argmax}\{\lambda_j(x_k) \mid j \in S_k\} \subseteq J_k$ and $d_k = x_k - e_{\hat\jmath}$ with $\hat\jmath \in J_k$. In particular, by (3.13) we get

$$\max\{\lambda_j(x_k) \mid j \in J_k\} = \lambda_{\hat\jmath}(x_k) > \frac{\delta_k}{2} \,. \tag{3.15}$$

We now want to show that $\alpha_k = \alpha_k^{\max}$. Assume by contradiction $\alpha_k < \alpha_{\max}$. Then by the lower bound on the stepsize and (3.13)

$$\alpha_k \geq \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2} = \frac{\lambda_i(x_k)}{L\|d_k\|^2} \geq \frac{\delta_{\min}}{2L\|d_k\|^2} \,, \tag{3.16}$$

where in the last inequality we used (3.15) together with $\delta_k \geq \delta_{\min}$. Also, by Lemma 7.1

$$\begin{aligned} &\|d_k\| = \|e_{\hat\jmath} - x_k\| \leq \sqrt{2}(e_{\hat\jmath} - x_k)_{\hat\jmath} = -\sqrt{2}(d_k)_{\hat\jmath} \Rightarrow \frac{(d_k)_{\hat\jmath}}{\|d_k\|^2} \leq \frac{(d_k)_{\hat\jmath}}{\|d_k\|\sqrt{2}} \leq -1/2 \\ &(x_k)_{\hat\jmath} = (x_k - x^*)_{\hat\jmath} \leq \frac{\|x_k - x^*\|_1}{2} < \frac{r_*}{2} = \frac{\delta_{\min}}{4L + 2\delta_{\min}}. \end{aligned} \tag{3.17}$$

Finally, combining (3.17) with (3.16)

$$\begin{aligned} (x_{k+1})_{\hat\jmath} &= (x_k)_{\hat\jmath} + (d_k)_{\hat\jmath}\alpha_k < \frac{r_*}{2} - \frac{\|d_k\|^2}{2}\alpha_k \leq \frac{r_*}{2} - \frac{\|d_k\|^2}{2}\frac{\delta_{\min}}{2L\|d_k\|^2} \\ &= \frac{\delta_{\min}}{4L + 2\delta_{\min}} - \frac{\delta_{\min}}{4L} < 0, \end{aligned}$$

where we used (3.16) to bound $\alpha_k$ in the first inequality, (3.17) to bound $(x_k)_{\hat\jmath}$ and $\frac{(d_k)_{\hat\jmath}}{\|d_k\|^2}$. Hence $(x_{k+1})_{\hat\jmath} < 0$, contradiction. ∎

**4. Active set complexity bounds.** Before giving the active set complexity bounds in several settings it is important to clarify that by active set associated to a stationary point $x^*$ we do not mean the set $\operatorname{supp}(x^*)^c = \{i \in [1:n] \mid (x^*)_i = 0\}\}$ but the set $I^c(x^*) = \{i \in [1:n] \mid \lambda_i(x^*) > 0\}$. In general $I^c(x^*) \subset \operatorname{supp}(x^*)^c$ by complementarity conditions, with

$$\operatorname{supp}(x^*)^c = I^c(x^*) \Leftrightarrow \text{complementarity conditions are strict in } x^*. \tag{4.1}$$

The face $\mathcal{F}$ of $\Delta_{n-1}$ defined by the constraints with indices in $I^c(x^*)$ still has a nice geometrical interpretation: it is the face of $\Delta_{n-1}$ exposed by $-\nabla f(x^*)$.
It is at this point natural to require that the sequence $\{x_k\}$ converges to a subset $A$ of $\mathcal{X}^*$ for which $I^c$ is constant. This motivates the following definition:

DEFINITION 4.1. *A compact subset $A$ of $\mathcal{X}^*$ is said to have the* support identification property (SIP) *if there exists an index set $I_A^c \subset [1:n]$ such that*

$$I^c(x) = I_A^c \quad \textit{for all } x \in A \,.$$

The geometrical interpretation of the above definition is the following: for every point in the subset $A$ the negative gradient $-\nabla f(x^*)$ exposes the same face. This is trivially true if $A$ is a singleton, and it is also true if for instance $A$ is contained in the relative interior of a face of $\Delta_{n-1}$ and strict complementarity conditions hold for every point in this face. We further define

$$\delta_{\min}(A) = \min\{\lambda_i(x) \mid x \in A,\ i \in I_A^c\}\ .$$

Notice that by the compactness of $A$ we always have $\delta_{\min}(A) > 0$ if $A$ enjoys the SIP. We can finally give a rigorous definition of what it means to solve the active set problem:

DEFINITION 4.2. *Consider an algorithm generating a sequence $\{x_k\}$ converging to a subset $A$ of $\mathcal{X}^*$ enjoying the SIP. We will say that this algorithm solves the active set problem in $M$ steps if $(x_k)_i = 0$ for every $i \in I_A^c$, $k \geq M$.*

We can now apply Theorem 3.1 to show that once a sequence is definitely close enough to a set enjoying the SIP, the AFW identifies the active set in at most $|I^c|$ steps. We first need to define a quantity that we will use as a lower bound on the stepsizes:

$$\bar{\alpha}_k = \min\left(\alpha_k^{\max}, \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}\right)\ , \tag{4.2}$$

THEOREM 4.1. *Let $\{x_k\}$ be a sequence generated by the AFW, with stepsize $\alpha_k \geq \bar{\alpha}_k$. Let $\mathcal{X}^*$ be the set of stationary points of a function $f : \Delta_{n-1} \to \mathbb{R}$ with $\nabla f$ having Lipschitz constant $L$. Assume that there exists a compact subset $A$ of $\mathcal{X}^*$ with the SIP such that $x_k \to A$. Then there exists $M$ such that*

$$(x_k)_i = 0 \quad \text{for every } k \geq M \text{ and all } i \in I_A^c\,.$$

*Proof.* Let $J_k = \{i \in I_A^c \mid (x_k)_i > 0\}$ and choose $\bar{k}$ such that $\operatorname{dist}_1(x_k, A) < \frac{\delta_{\min}(A)}{2L+\delta_{\min}(A)} = r_*$ for every $k \geq \bar{k}$. Then for every $k \geq \bar{k}$ there exists $y^* \in A$ with $\|x_k - y^*\|_1 < r_*$. But since by hypothesis for every $y^* \in A$ the support of the multiplier function is $I_A^c$ with $\delta_{\min}(A) \leq \lambda_i(y^*)$ for every $i \in I_A^c$, we can apply Theorem 3.1 with $y^*$ as fixed point and obtain that $|J_{k+1}| \leq \max(0, |J_k| - 1)$. This means that it takes at most $|J_{\bar{k}}| \leq |I_A^c|$ steps for all the variables with indices in $I_A^c$ to be 0. Again by (3.8), we conclude by induction $|J_k| = 0$ for every $k \geq M = \bar{k} + |I_A^c|$, since $|J_{\bar{k}+|I_A^c|}| = 0$. ∎

The proof above also gives a relatively simple upper bound for the complexity of the active set problem:

PROPOSITION 4.1. *Under the assumptions of Theorem 4.1, the active set complexity is at most*

$$\min\{\bar{k} \in \mathbb{N}_0 \mid \operatorname{dist}_1(x_k, A) < r_* \forall k \geq \bar{k}\} + |I_A^c|,$$

*where $r_* = \frac{\delta_{\min}(A)}{2L+\delta_{\min}(A)}$.*

We now report an explicit bound for the strongly convex case, and analyze in depth the nonconvex case in Section 5. From strong convexity of $f$, it is easy to see that the following inequality holds for every $x$ on $\Delta_{n-1}$:

$$f(x) \geq f(x^*) + \frac{u_1}{2}\|x - x^*\|_1^2, \tag{4.3}$$

with $u_1 > 0$.

COROLLARY 4.1. *Let $\{x_k\}$ be the sequence of points generated by AFW with $\alpha_k \geq \bar{\alpha}_k$. Assume that $f$ is strongly convex and let*

$$h_k \leq q^k h_0, \tag{4.4}$$

*with $q < 1$ and $h_k = f(x_k) - f_*$, be the convergence rate related to AFW. Then the active set complexity is*

$$\max\left(0, \left\lceil \frac{\ln(h_0) - \ln(u_1 r_*^2/2)}{\ln(1/q)} \right\rceil\right) + |I^c|\ .$$

*Proof.* Notice that by the linear convergence rate (4.4), and the fact that $q < 1$, the number of steps needed to reach the condition

$$h_k \leq \frac{u_1}{2} r_*^2 \tag{4.5}$$

is at most

$$\bar{k} = \max\left(0, \left\lceil \frac{\ln(h_0) - \ln(u_1 r_*^2/2)}{\ln(1/q)} \right\rceil\right) .$$

We claim that if condition (4.5) holds then it takes at most $|I^c|$ steps for the sequence to be definitely in the active set.
Indeed if $q^k h_0 \leq \frac{u_1}{2} r_*^2$ then necessarily $x_k \in B_1(x^*, r_*)$ by (4.3), and by monotonicity of the bound (4.4) we then have $x_{k+h} \in B_1(x^*, r_*)$ for every $h \geq 0$. Once the sequence is definitely in $B_1(x^*, r_*)$ by (3.8) it takes at most $|J_{\bar{k}}| \leq |I^c|$ steps for all the variables with indices in $I^c$ to be 0. To conclude, again by (3.8) since $|J_{\bar{k}+|I^c|}| = 0$ by induction $|J_m| = 0$ for every $m \geq \bar{k} + |I^c|$. ∎

Remark 4.1. *We would like to notice that strong convexity of $f$ in Corollary 4.1 might actually be replaced by condition given in (4.3) if we assume the linear rate (4.4) (which may not hold in the nonconvex case).*

The proof of AFW active set complexity for generic polytopes in the strongly convex case requires additional theoretical results and is presented in the appendix.

**5. Active set complexity for nonconvex objectives.** In this section, we focus on problems with nonconvex objectives. We first give a more explicit convergence rate for AFW in the nonconvex case, then we prove a general active set identification result for the method. Finally, we analyze both local and global active set complexity bounds related to AFW. A fundamental element in our analysis will be the FW gap function $g : \Delta_{n-1} \to \mathbb{R}$ defined as

$$g(x) = \max_{i \in [1:n]} \{-\lambda_i(x)\} .$$

We clearly have $g(x) \geq 0$ for every $x \in \Delta_{n-1}$, with equality iff $x$ is a stationary point. The reason why this function is called FW gap is evident from the relation

$$g(x_k) = -\nabla f(x_k)^\top d_k^{\mathcal{FW}}.$$

This is a standard quantity appearing in the analysis of FW variants (see, e.g., [25] ) and is computed for free at each iteration of a FW-like algorithm. In [27], the author uses the gap to analyze the convergence rate of the classic FW algorithm in the nonconvex case. More specifically, a convergence rate of $O(\frac{1}{\sqrt{k}})$ is proved for the minimal FW gap up to iteration $k$:

$$g_k^* = \min_{0 \leq i \leq k-1} g(x_i).$$

The results extend in a nice and straightforward way the ones reported in [29] for proving the convergence of gradient methods in the nonconvex case. Inspired by the analysis of the AFW method for strongly convex objectives reported in [32], we now study the AFW convergence rate in the nonconvex case with respect to the sequence $\{g_k^*\}$.

In the rest of this section we assume that the AFW starts from a vertex of the probability simplex. This is not a restrictive assumption. By exploiting affine invariance one can indeed apply the same theorems to the AFW starting from $e_{n+1}$ for $\tilde{f} : \Delta_n \to \mathbb{R}$ satisfying

$$\tilde{f}(y) = f(y_1 e_1 + ... y_n e_n + y_{n+1} p),$$

where $p \in \Delta_{n-1}$ is the desired starting point. We will discuss more in detail the invariance of the AFW under affine transformations in Section 7.2.

**5.1. Global convergence.** We start investigating the minimal FW gap, giving estimates of rates of convergence:

THEOREM 5.1. *Let $f^* = \min_{x\in\Delta_{n-1}} f(x)$, and let $\{x_k\}$ be a sequence generated by the AFW algorithm applied to $f$ on $\Delta_{n-1}$, with $x_0$ a vertex of $\Delta_{n-1}$. Assume that the stepsize $\alpha_k$ is larger or equal than $\bar\alpha_k$ (as defined in (4.2)), and that*

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \rho\bar\alpha_k \left(-\nabla f(x_k)^\top d_k\right) \tag{5.1}$$

*for some fixed $\rho > 0$. Then for every $T \in \mathbb{N}$*

$$g_T^* \leq \max\left(\sqrt{\frac{4L(f(x_0)-f^*)}{\rho T}}, \frac{4(f(x_0)-f^*)}{T}\right) .$$

*Proof.* Let $r_k = -\nabla f(x_k)$ and $g_k = g(x_k)$. We distinguish three cases.

**Case 1.** $\bar\alpha_k < \alpha_k^{\max}$. Then $\bar\alpha_k = \frac{-\nabla f(x_k)^\top d_k}{L\|d_k\|^2}$ and relation (5.1) becomes

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \rho\bar\alpha_k r_k^\top d_k = \frac{\rho}{L\|d_k\|^2}(r_k^\top d_k)^2$$

and consequently

$$f(x_k) - f(x_{k+1}) \geq \frac{\rho}{L\|d_k\|^2}(r_k^\top d_k)^2 \geq \frac{\rho}{L\|d_k\|^2} g_k^2 \geq \frac{\rho g_k^2}{2L}, \tag{5.2}$$

where we used $r_k^\top d_k \geq g_k$ in the second inequality and $\|d_k\| \leq \sqrt{2}$ in the third one.
As for $S_k$, by hypothesis we have either $d_k = d_k^{\mathcal{FW}}$ so that $d_k = e_i - x_k$ or $d_k = d_k^{\mathcal{A}} = x_k - e_i$ for some $i \in \{1, ..., n\}$. In particular $S_{k+1} \subseteq S_k \cup \{i\}$ so that $|S_{k+1}| \leq |S_k| + 1$.
**Case 2:** $\alpha_k = \bar\alpha_k = \alpha_k^{\max} = 1, d_k = d_k^{\mathcal{FW}}$. By the standard descent lemma [5, Proposition 6.1.2] applied to $f$ with center $x_k$ and $\alpha = 1$

$$f(x_{k+1}) = f(x_k + d_k) \leq f(x_k) + \nabla f(x_k)^\top d_k + \frac{L}{2}\|d_k\|^2 .$$

Since by the Case 2 condition $\min\left(\frac{-\nabla f(x_k)^\top d_k}{\|d_k\|^2 L}, 1\right) = \alpha_k = 1$ we have

$$\frac{-\nabla f(x_k)^\top d_k}{\|d_k\|^2 L} \geq 1 \text{ , so } \quad -L\|d_k\|^2 \geq \nabla f(x_k)^\top d_k ,$$

hence we can write

$$f(x_k) - f(x_{k+1}) \geq -\nabla f(x_k)^\top d_k - \frac{L}{2}\|d_k\|^2 \geq -\frac{\nabla f(x_k)^\top d_k}{2} \geq \frac{1}{2} g_k . \tag{5.3}$$

Reasoning as in Case 1 we also have $|S_{k+1}| \leq |S_k| + 1$.
**Case 3:** $\alpha_k = \bar\alpha_k = \alpha_k^{\max}$, $d_k = d_k^{\mathcal{A}}$. Then $d_k = x_k - e_i$ for $i \in S_k$ and

$$(x_{k+1})_j = (1 + \alpha_k)(x_k)_j - \alpha_k (e_i)_j,$$

with $\alpha_k = \alpha_k^{\max} = \frac{(x_k)_i}{1-(x_k)_i}$. Therefore $(x_{k+1})_j = 0$ for $j \in \{1, ..., n\} \setminus S_k \cup \{i\}$ and $(x_{k+1})_j \neq 0$ for $j \in S_k \setminus \{i\}$. In particular $|S_{k+1}| = |S_k| - 1$.

For $i = 1, 2, 3$ let now $n_i(T)$ be the number of Case $i$ steps done in the first $T$ iterations of the AFW. We have by induction on the recurrence relation we proved for $|S_k|$

$$|S_T| - |S_0| \leq n_1(T) + n_2(T) - n_3(T) , \tag{5.4}$$

for every $T \in \mathbb{N}$.
Since $n_3(T) = T - n_1(T) - n_2(T)$ from (5.4) we get

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T}{2} \ ,$$

where we used $|S_0| = 1 \leq |S_T|$. Let now $C_i^T$ be the set of iteration counters up to $T-1$ corresponding to Case $i$ steps for $i \in \{1,2,3\}$, which satisfies $|C_i^T| = n_i(T)$. We have by summing (5.2) and (5.3) for the indices in $C_1^T$ and $C_2^T$ respectively

$$\sum_{k \in C_1^T} f(x_k) - f(x_{k+1}) + \sum_{k \in C_2^T} f(x_{k+1}) - f(x_k) \geq \sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \ . \tag{5.5}$$

We now lower bound the right-hand side of (5.5) in terms of $g_T^*$ as follows:

$$\begin{aligned}
&\sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \geq |C_1^T| \min_{k \in C_1^T} \frac{\rho g_k^2}{2L} + |C_2^T| \min_{k \in C_2^T} \frac{g_k}{2} \geq \\
\geq &(|C_1^T| + |C_2^T|) \min\left( \frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) = [n_1(T) + n_2(T)] \min\left( \rho \frac{(g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) \geq \\
\geq &\frac{T}{2} \min\left( \frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) \ .
\end{aligned}$$

Since the left-hand side of (5.5) can clearly be upper bounded by $f(x_0) - f^*$ we have

$$f(x_0) - f^* \geq \frac{T}{2} \min\left( \frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) \ .$$

To finish, if $\frac{T}{2} \min\left( \frac{g_T^*}{2}, \frac{\rho (g_T^*)^2}{2L} \right) = \frac{T g_T^*}{4}$ we then have

$$g_T^* \leq \frac{4(f(x_0) - f^*)}{T} \tag{5.6}$$

and otherwise

$$g_T^* \leq \sqrt{\frac{4L(f(x_0) - f^*)}{\rho T}} \ . \tag{5.7}$$

The claim follows by taking the max in the system formed by (5.6) and (5.7). ∎

When the stepsizes coincide with the lower bounds $\bar{\alpha}_k$ or are obtained using exact linesearch, we have the following corollary:

Corollary 5.1. *Under the assumptions of Theorem 5.1, if $\alpha_k = \bar{\alpha}_k$ or if $\alpha_k$ is selected by exact linesearch then for every $T \in \mathbb{N}$*

$$g_T^* \leq \max\left( \sqrt{\frac{8L(f(x_0) - f^*)}{T}}, \frac{4(f(x_0) - f^*)}{T} \right) \ . \tag{5.8}$$

*Proof.* By points 2 and 3 of Lemma 7.2, relation (5.1) is satisfied with $\rho = \frac{1}{2}$ for both $\alpha_k = \bar{\alpha}_k$ and $\alpha_k$ given by exact linesearch, and we also have $\alpha_k \geq \bar{\alpha}_k$ in both cases. The conclusion follows directly from Theorem 5.1. ∎

**5.2. A general active set identification result.** We can now give a general active set identification result in the nonconvex setting. While we won't use strict complementarity when the stepsizes are given by (4.2), without this assumption we will need strict complementarity. Notice that if $A \subseteq \mathcal{X}^*$ enjoys the SIP and if strict complementarity is satisfied for every $x \in A$, then as a direct consequence of (4.1) we have

$$\text{supp}(x) = [1:n] \setminus I^c(x) = [1:n] \setminus I^c_A \tag{5.9}$$

for every $x \in A$. In this case we can then define $\text{supp}(A)$ as the (common) support of the points in $A$.

THEOREM 5.2. *Let $\{x_k\}$ be the sequence generated by the AFW method with stepsizes satisfying $\alpha_k \geq \bar{\alpha}_k$ and (5.1), where $\bar{\alpha}_k$ is given by (4.2). Let $\mathcal{X}^*$ be the subset of stationary points of $f$. We have:*

*(a) $x_k \to \mathcal{X}^*$.*

*(b) If $\alpha_k = \bar{\alpha}_k$ then $\{x_k\}$ converges to a connected component $A$ of $\mathcal{X}^*$. If additionally $A$ has the SIP then $\{x_k\}$ identifies $I^c_A$ in finite time.*

*Assume now that $\mathcal{X}^* = \bigcup_{i=1}^C A_i$ with $\{A_i\}_{i=1}^C$ compact, with distinct supports and such that $A_i$ has the SIP for each $i \in [1\!:\!C]$.*

*(c) If $\alpha_k \geq \bar{\alpha}_k$ and if strict complementarity holds for all points in $\mathcal{X}^*$ then $\{x_k\}$ converges to $A_l$ for some $l \in [1:C]$ and identifies $I^c_{A_l}$ in finite time.*

*Proof.* a) By the proof of Theorem 5.1 and the continuity of the multiplier function we have

$$x_{k(j)} \to g^{-1}(0) = \mathcal{X}^* \ , \tag{5.10}$$

where $\{k(j)\}$ is the sequence of indexes corresponding to Case 1 or Case 2 steps. Let $k'(j)$ be the sequence of indexes corresponding to Case 3 steps. Since for such steps $\alpha_{k'(j)} = \bar{\alpha}_{k'(j)}$ we can apply Corollary 7.1 to obtain

$$\|x_{k'(j)} - x_{k'(j)+1}\| \to 0 \ . \tag{5.11}$$

Combining (5.10), (5.11) and the fact that there can be at most $n-1$ consecutive Case 3 steps, we get $x_k \to \mathcal{X}^*$.

b) By the boundedness of $f$ and point 2 of Lemma 7.2 if $\alpha_k = \bar{\alpha}_k$ then $\|x_{k+1} - x_k\| \to 0$. It is a basic topology fact that if $\{x_k\}$ is bounded and $\|x_{k+1} - x_k\| \to 0$ then the set of limit points of $\{x_k\}$ is connected. This together with point a) ensures that the set of limit points must be contained in a connected component $A$ of $\mathcal{X}^*$. By Theorem 4.1 it follows that if $A$ has constant support $\{x_k\}$ identifies $I^c_A$ in finite time.

c) Consider a disjoint family of subsets $\{U_i\}_{i=1}^C$ of $\Delta_{n-1}$ with $U_i = \{x \in \Delta_{n-1} \mid \text{dist}_1(x, A_i) \leq r_i\}$ where $r_i$ is small enough to ensure some conditions that we now specify. First, we need

$$r_i < \frac{\delta_{\min}(A_i)}{2L + \delta_{\min}(A_i)}$$

so that $r_i$ is smaller than the active set radius of every $x \in A_i$ and in particular for every $x \in U_i$ there exists $x^* \in A_i$ such that

$$\|x - x^*\|_1 < \frac{\delta_{\min}(x^*)}{2L + \delta_{\min}(x^*)}. \tag{5.12}$$

Second, we choose $r_i$ small enough so that $\{U_i\}_{i=1}^C$ are disjoint and

$$\text{supp}(y) \supseteq \text{supp}(A_i) \ \forall y \in U_i \ , \tag{5.13}$$

where these conditions can be always satisfied thanks to the compactness of $A_i$.

Assume now by contradiction that the set $S$ of limit points of $\{x_k\}$ intersects more than one of the $\{A_i\}_{i=1}^C$. Let in particular $A_l$ minimize $|\text{supp}(A_l)|$ among the sets containing points of $S$. By point a)

$x_k \in \cup_{i=1}^C U_i$ for $k \geq M$ large enough and we can define an infinite sequence $\{t(j)\}$ of exit times greater than $M$ for $U_l$ so that $x_{t(j)} \in U_l$ and $x_{t(j)+1} \in \cup_{i\in[1:C]\setminus l} U_i$. Up to considering a subsequence we can assume $x_{t(j)+1} \in U_m$ for a fixed $m \neq l$ for every $j \in \mathbb{N}_0$.
We now distinguish two cases as in the proof of Theorem 3.1, where notice that by equation (5.12) the hypotheses of Theorem 3.1 are satisfied for $k = t(j)$ and some $x^* \in A_l$.
**Case 1.** $(x_{t(j)})_h = 0$ for every $h \in I^c_{A_l}$. In the notation of Theorem 3.1 this corresponds to the case $|J_{t(j)}| = 0$. Then by (3.10) we also have $\lambda_h(x_{t(j)}) > 0$ for every $h \in I^c_{A_l}$. Thus $(x_{t(j)+1})_h = (x_{t(j)})_h = 0$ for every $h \in I^c_{A_l}$ by Lemma 3.2, so that we can write

$$\operatorname{supp}(A_m) \subseteq \operatorname{supp}(x_{t(j)+1}) \subseteq [1:n] \setminus I^c_{A_l} = \operatorname{supp}(A_l), \tag{5.14}$$

where the first inclusion is justified by (5.13) for $i = m$ and the second by strict complementarity (see also (5.9) and the related discussion). But since by hypothesis $\operatorname{supp}(A_m) \neq \operatorname{supp}(A_l)$ the inclusion (5.14) is strict and so it is in contradiction with the minimality of $|\operatorname{supp}(A_l)|$.
**Case 2.** $|J_{t(j)}| > 0$. Then reasoning as in the proof of Theorem 3.1 we obtain $d_{t(j)} = x_{t(j)} - e_{\bar h}$ for some $\bar h \in J_{t(j)} \subset I^c_{A_l}$. Let $\tilde x^* \in A_l$, and let $\tilde d = \alpha_{t(j)} d_{t(j)}$. The sum of the components of $\tilde d$ is 0 with the only negative component being $\tilde d_{\bar h}$ and therefore

$$\tilde d_{\bar h} = -\sum_{h\in[1:n]\setminus\bar h} \tilde d_h = -\sum_{h\in[1:n]\setminus\bar h} |\tilde d_h| \tag{5.15}$$

We claim that $\|x_{t(j)+1} - \tilde x^*\|_1 \leq \|x_{t(j)} - \tilde x^*\|_1$. This is enough to finish because since $\tilde x^* \in A_l$ is arbitrary then it follows $\operatorname{dist}_1(x_{t(j)+1}, A_l) \leq \operatorname{dist}_1(x_{t(j)}, A_l)$ so that $x_{t(j)+1} \in U_l$, a contradiction.
We have

$$\begin{aligned}
&\|\tilde x^* - x_{t(j)+1}\|_1 = \|\tilde x^* - x_{t(j)} - \alpha_{t(j)} d_{t(j)}\|_1 = \\
=&|\tilde x^*_{\bar h} - (x_{t(j)})_{\bar h} - \tilde d_{\bar h}| + \sum_{h\in[1:n]\setminus\bar h} |\tilde x^*_h - (x_{t(j)})_h - \tilde d_h| = \\
=&|\tilde x^*_{\bar h} - (x_{t(j)})_{\bar h}| + \tilde d_{\bar h} + \sum_{h\in[1:n]\setminus\bar h} |\tilde x^*_h - (x_{t(j)})_h - \tilde d_h| \leq \\
\leq&|\tilde x^*_{\bar h} - (x_{t(j)})_{\bar h}| + \tilde d_{\bar h} + \sum_{h\in[1:n]\setminus\bar h} (|\tilde x^*_h - (x_{t(j)})_h| + |\tilde d_h|) = \\
=&\|x_{t(j)} - \tilde x^*\|_1 + \tilde d_{\bar h} + \sum_{h\in[1:n]\setminus\bar h} |\tilde d_h| = \|x_{t(j)} - \tilde x^*\|_1
\end{aligned}$$

where in the third equality we used $0 = \tilde x^*_{\bar h} \leq -\tilde d_{\bar h} \leq (x_{t(j)})_{\bar h}$ and in the last equality we used (5.15). Reasoning by contradiction we have proved that all the limit points of $\{x_k\}$ are in $A_l$ for some $l \in [1, ..., C]$. The conclusion follows immediately from Theorem 4.1. ∎

**5.3. Quantitative version of active set identification.** We now assume that the gap function $g(x)$ satisfies the Hölderian error bound condition

$$g(x) \geq \theta \operatorname{dist}_1(x, \mathcal{X}^*)^p \tag{5.16}$$

for some $\theta, p > 0$ (see e.g. [7] for some example). This is true for instance if the components of $\nabla f(x)$ are semialgebraic functions. We have the following active set complexity bound:

THEOREM 5.3. *Assume $\mathcal{X}^* = \bigcup_{i\in[1:C]} A_i$ where $A_i$ is compact and with the SIP for every $i \in [1:C]$ and $0 < d \stackrel{\text{def}}{=} \min_{\{i,j\}\subset[1:C]} \operatorname{dist}_1(A_i, A_j)$. Let $r_*$ be the minimum active set radius of the sets $\{A_i\}_{i=1}^C$. Let $q(\varepsilon) : \mathbb{R}_{>0} \to \mathbb{N}_0$ be such that $f(x_k) - f(x_{k+1}) \leq \varepsilon$ for every $k \geq q(\varepsilon)$, and assume that $g(x)$ satisfies (5.16). Assume that the stepsizes satisfy $\alpha_k = \bar\alpha_k$, with $\bar\alpha_k$ given by (4.2). Then the active set complexity*

*is at most* $q(\bar\varepsilon)+2n$ *for* $\bar\varepsilon$ *satisfying the following conditions*

$$\bar\varepsilon < L\,, \quad \left(\frac{2\sqrt{L\bar\varepsilon}}{\theta}\right)^{\frac{1}{p}} < r_* \quad \textit{and} \quad 2\left(\frac{2\sqrt{L\bar\varepsilon}}{\theta}\right)^{\frac{1}{p}} + 2n\sqrt{\frac{2\bar\varepsilon}{L}} \leq d\,. \tag{5.17}$$

The proof is substantially a quantitative version of the argument used to prove point b) of Theorem 5.2.

*Proof.* Fix $k \geq q(\bar\varepsilon)$, so that

$$f(x_k) - f(x_{k+1}) \leq \bar\varepsilon\,. \tag{5.18}$$

We will refer to Case $i$ steps for $i \in [1:3]$ following the definitions in Theorem 5.1. If the step $k$ is a Case 1 step, then by (5.2) with $\rho = 1/2$ we have

$$f(x_k) - f(x_{k+1}) \geq \frac{g(x_k)^2}{4L}$$

and this together with (5.18) implies

$$2\sqrt{L\bar\varepsilon} \geq 2\sqrt{L(f(x_k) - f(x_{k+1}))} \geq g(x_k)\,.$$

Analogously, if the step $k$ is a Case 2 step, then by (5.3) we have

$$f(x_k) - f(x_{k+1}) \geq \frac{g(x_k)}{2}$$

so that $2\bar\varepsilon \geq g(x_k)$. By the leftmost condition in (5.17) we have $\bar\varepsilon < L$ so that $2\sqrt{L\bar\varepsilon} \geq 2\bar\varepsilon$, and therefore for both Case 1 and Case 2 steps we have

$$g(x_k) \leq 2\sqrt{L\bar\varepsilon}\,. \tag{5.19}$$

By inverting relation (7.1), we also have

$$\|x_k - x_{k+1}\| \leq \sqrt{\frac{2(f(x_k) - f(x_{k+1}))}{L}} \leq \sqrt{\frac{2\bar\varepsilon}{L}}\,. \tag{5.20}$$

Now let $\bar k \geq q(\bar\varepsilon)$ be such that step $\bar k$ is a Case 1 or Case 2 step. By the error bound condition together with (5.19)

$$\operatorname{dist}_1(x_{\bar k}, \mathcal{X}^*) \leq \left(\frac{g(x_{\bar k})}{\theta}\right)^{\frac{1}{p}} \leq \left(\frac{2\sqrt{L\bar\varepsilon}}{\theta}\right)^{\frac{1}{p}} < r_*\,, \tag{5.21}$$

where we used (5.19) in the second inequality and the second condition of (5.17) in the third inequality. In particular there exists $l$ such that $\operatorname{dist}_1(x_{\bar k}, A_l) \leq (2\sqrt{L\bar\varepsilon}/\theta)^{1/p}$. We claim now that $I^c_{A_l}$ is identified at latest at step $\bar k + n$.

First, we claim that for every Case 1 or Case 2 step with index $\tau \geq \bar k$ we have $\operatorname{dist}_1(x_\tau, A_l) \leq (g(x_\tau)/\theta)^{1/p}$. We reason by induction on the sequence $\{s(k')\}$ of Case 1 or Case 2 steps following $\bar k$, so that in particular $s(1) = \bar k$ and $\operatorname{dist}_1(x_{s(1)}, A_l) \leq g(x_{s(1)})$ is true by (5.21). Since there can be at most $n-1$ consecutive Case 3 steps, we have $s(k'+1) - s(k') \leq n$ for every $k' \in \mathbb{N}_0$. Therefore

$$\begin{aligned}\|x_{s(k')} - x_{s(k'+1)}\|_1 \leq \sum_{i=s(k')}^{s(k'+1)-1} \|x_{i+1} - x_i\|_1 \leq 2 \sum_{i=s(k')}^{s(k'+1)-1} \|x_{i+1} - x_i\| \leq \\ \leq 2[s(k'+1) - s(k')]\sqrt{\frac{2\bar\varepsilon}{L}} \leq 2n\sqrt{\frac{2\bar\varepsilon}{L}}\,,\end{aligned} \tag{5.22}$$

where in the second inequality we used part 3 of Lemma 7.1 to bound each of the summands of the left-hand side, and in the third inequality we used (5.20). Assume now by contradiction $\mathrm{dist}_1(x_{s(k'+1)}, A_l) > (g(x_{s(k'+1)})/\theta)^{1/p}$. Then by (5.21) applied to $s(k'+1)$ instead of $\bar{k}$ there must exists necessarily $j \neq l$ such that $\mathrm{dist}_1(x_{s(k'+1)}, A_j) \leq (g(x_{s(k'+1)})/\theta)^{1/p}$. In particular we have

$$\begin{aligned} \|x_{s(k')} - x_{s(k'+1)}\|_1 \geq & \mathrm{dist}_1(A_l, A_j) - \mathrm{dist}_1(x_{s(k'+1)}, A_j) - \mathrm{dist}_1(x_{s(k')}, A_l) \geq \\ \geq & d - \left(\frac{g(x_{s(k')})}{\theta}\right)^{\frac{1}{p}} - \left(\frac{g(x_{s(k'+1)})}{\theta}\right)^{\frac{1}{p}} \geq d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} , \end{aligned} \tag{5.23}$$

where we used (5.19) in the last inequality. But by the second condition of (5.17), we have

$$d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} > 2n\sqrt{\frac{2\bar{\varepsilon}}{L}} . \tag{5.24}$$

Concatenating (5.22), (5.24) and (5.23) we get a contradiction and the claim is proved. Notice that an immediate consequence of this claim is $\mathrm{dist}_1(x_\tau, A_l) < r_*$ by (5.21) applied to $\tau$ instead of $\bar{k}$, where $\tau \geq \bar{k}$ is an index corresponding to a Case 1 or Case 2 step.
To finish the proof, first notice that there exists an index $\bar{k} \in [q(\bar{\varepsilon}), q(\bar{\varepsilon}) + n]$ corresponding to a Case 1 or Case 2 step, since there can be at most $n-1$ consecutive Case 3 steps. Furthermore, since by (5.21) we have $\mathrm{dist}_1(x_{\bar{k}}, A_l) < r_*$, by the local identification Theorem 3.1 in the steps immediately after $\bar{k}$ the AFW identifies one at a time the variables in $I^c_{A_l}$, so that there exists $h \leq n$ such that $(x_{\bar{k}+h})_i = 0$ for every $i \in I^c_{A_l}$. Moreover, by the claim every Case 1 and Case 2 step following step $\bar{k}$ happens for points inside $B_1(A_l, r_*)$ so it does not change the components corresponding to $I^c_{A_l}$ by the local identification Theorem 3.1. At the same time, Case 3 steps do not increase the support, so that $(x_{\bar{k}+h+l})_i = 0$ for every $i \in I^c_{A_l}$, $l \geq 0$. Thus active set identification happens in $\bar{k} + h \leq q(\bar{\varepsilon}) + n + h \leq q(\bar{\varepsilon}) + 2n$ steps. □

Remark 5.1. *Assume that the set of stationary points is finite, so that $A_i = \{a_i\}$ for every $i \in [1{:}C]$ with $a_i \in \Delta_{n-1}$. Let*

$$c_{\min} = \min_{i \in [1:C]} \min_{j:(a_i)_j \neq 0} (a_i)_j \tag{5.25}$$

*be the minimal nonzero component of a stationary point. Then one can prove a $q(\bar{\varepsilon}) + n$ active set identification bound replacing (5.17) with the following condition on $\bar{\varepsilon}$ which has no explicit dependence on $n$:*

$$\bar{\varepsilon} < L, \quad r(\bar{\varepsilon}) + l(\bar{\varepsilon}) < \min(r_*, d/2, c_{\min}/2) ,$$

*where $r(\bar{\varepsilon}) = \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}}$ and $l(\bar{\varepsilon}) = 2\sqrt{\frac{2\bar{\varepsilon}}{L}}$. We do not discuss the proof since it roughly follows the same lines of Theorem 5.3's proof.*

Remark 5.2. *When we have an explicit expression for the convergence rate $q(\varepsilon)$, then we can get an active set complexity bound using Theorem 5.3.*

**5.4. Local active set complexity bound.** A key element to ensure local convergence to a strict local minimum will be the following property

$$x_k \in \mathrm{argmax}\{f(x) \mid x \in \mathrm{conv}(x_k, x_{k+1})\} . \tag{5.26}$$

which in particular holds when $\alpha_k = \bar{\alpha}_k$ as it is proved in Lemma 7.2. The property (5.26) is obviously stronger than the usual monotonicity, and it ensures that the sequence cannot escape from connected components of sublevel sets. When $f$ is convex it is immediate to check that (5.26) holds if and only if $\{f(x_k)\}$ is monotone non increasing.

Let $x^*$ be a stationary point which is also a strict local minimizer isolated from the other stationary points and $\tilde{f} = f(x^*)$. Let then $\beta$ be such that there exists a connected component $V_{x^*,\beta}$ of $f^{-1}((-\infty, \beta])$ satisfying

$$V_{x^*,\beta} \cap \mathcal{X}^* = \{x^*\} = \operatorname{argmin}_{x \in V_{x^*,\beta}} f(x).$$

THEOREM 5.4. *Let $\{x_k\}$ be a sequence generated by the AFW, with $x_0 \in V_{x^*,\beta}$ and with stepsize given by (4.2). Let*

$$r_* = \frac{\delta_{\min}(x^*)}{2L + \delta_{\min}(x^*)} \ .$$

*Then $x_k \to x^*$ and the sequence identifies the support in at most*

$$\left\lceil \max\left( \frac{4(f(x_0) - \tilde{f})}{\tau}, \frac{8L(f(x_0) - \tilde{f})}{\tau^2} \right) \right\rceil + 1 + |I^c(x^*)|$$

*steps with*

$$\tau = \min\{g(x) \mid x \in f^{-1}([m, +\infty)) \cap V_{x^*,\beta}\} \ ,$$

*where*

$$m = \min\{ \ f(x) \mid x \in V_{x^*,\beta} \setminus B_{r_*}(x^*)\} \ .$$

*Proof.* We have all the hypotheses to apply the bound given in Corollary 5.1 for $g_k^*$:

$$g_k^* \leq \max\left( \sqrt{\frac{8L(f(x_0) - f^*)}{k}}, \frac{4(f(x_0) - f^*)}{k} \right) \ .$$

It is straightforward to check that if

$$\bar{h} = \left\lceil \max\left( \frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2} \right) \right\rceil + 1$$

then

$$g_{\bar{h}}^* < \tau \ .$$

Therefore, by the definition of $\tau$ ,we get $f(x_{\bar{h}}) < m$. We claim that $x_h \in B_{r_*}(x^*)$ for every $h \geq \bar{h}$. Indeed by point 1 of Lemma 7.2 the condition $\alpha_k = \bar{\alpha}_k$ on the stepsizes imply that $\{x_k\}$ satisfies (5.26) and it can not leave connected components of level sets. Thus since $f(x_h) < m$ we have

$$x_h \in V_{x^*,\beta} \cap f^{-1}(-\infty, m) \subset B_{r_*}(x^*) \ ,$$

where the inclusion follows directly from the definition of $m$. We can then apply the local active set identification Theorem 3.1 to obtain an active set complexity of

$$\bar{h} + |I^c(x^*)| = \left\lceil \max\left( \frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2} \right) \right\rceil + 1 + |I^c(x^*)| \ ,$$

thus getting our result. ∎

**6. Conclusions.** We proved general results for the AFW finite time active set convergence problem, giving explicit bounds on the number of steps necessary to identify the support of a solution. As applications of these results we computed the active set complexity for strongly convex functions and nonconvex functions. Possible expansions of these results would be to adapt them for other FW variants and, more generally, to other first order methods. It also remains to be seen if these identification properties of the AFW can be extended to problems with nonlinear constraints.

**7. Appendix.** In several proofs we need some elementary inequalities concerning the euclidean norm $\|\cdot\|$ and the norm $\|\cdot\|_1$.

LEMMA 7.1. *Given $\{x, y\} \subset \Delta_{n-1}$, $i \in [1:n]$:*
1. $\|e_i - x\| \le \sqrt{2}(e_i - x)_i$;
2. $(y - x)_i \le \|y - x\|_1/2$
3. *If $\{x_k\}$ is a sequence generated on the probability simplex by the AFW then $\|x_{k+1} - x_k\|_1 \le 2\|x_{k+1} - x_k\|$ for every $k$.*

*Proof.* 1. $(e_i - x)_j = -x_j$ for $j \neq i$, $(e_i - x)_i = 1 - x_i = \sum_{j\neq i} x_j$. In particular

$$\|e_i - x\| = (\sum_{j\neq i} x_j^2 + (e_i - x)_i^2)^{\frac{1}{2}} \le ((\sum_{j\neq i} x_j)^2 + (1 - x_i)^2)^{\frac{1}{2}} = \sqrt{2}(\sum_{j\neq i} x_j) = \sqrt{2}(e_i - x)_i$$

2. Since $\sum_{j\in[1:n]} x_j = \sum_{j\in[1:n]} y_j$ so that $\sum(x - y)_j = 0$ we have

$$(y - x)_i = \sum_{j\neq i}(x - y)_j$$

and as a consequence

$$\|y - x\|_1 = \sum_{j\in[1:n]} |(y - x)_j| \ge (y - x)_i + \sum_{j\neq i}(x - y)_j = 2(y - x)_i \ .$$

3. We have $x_{k+1} - x_k = \alpha_k d_k$ with $d_k = \pm(e_i - x_k)$ for some $i \in [1:n]$. By homogeneity it suffices to prove $\|d_k\| \ge \frac{1}{2}\|d_k\|_1$. We have

$$\|d_k\| \ge 1 - (x_k)_i = \frac{1}{2}(1 - (x_k)_i + \sum_{j\neq i}(x_k)_j) = \frac{1}{2}\|d_k\|_1 \ ,$$

where in the first equality we used $\sum_{i=1}^n (x_k)_i = 1$ and in the second equality we used $0 \le x_k \le 1$. ∎

**7.1. Technical results related to stepsizes.** We now prove several properties related to the stepsize given in (4.2).

LEMMA 7.2. *Consider a sequence $\{x_k\}$ in $\Delta_{n-1}$ such that $x_{k+1} = x_k + \alpha_k d_k$ with $\alpha_k \in \mathbb{R}_{\ge 0}$, $d_k \in \mathbb{R}^n$. Let $\bar\alpha_k$ be defined as in (4.2), let $p_k = -\nabla f(x_k)^\top d_k$ and assume $p_k > 0$. Then:*
1. *If $0 \le \alpha_k \le 2p_k/(\|d_k\|^2 L)$, the sequence $\{x_k\}$ has the property (5.26).*
2. *If $\alpha_k = \bar\alpha_k$ then (5.1) is satisfied with $\rho = \frac{1}{2}$. Additionally, we have*

$$f(x_k) - f(x_{k+1}) \ge L\frac{\|x_{k+1} - x_k\|^2}{2} \ . \tag{7.1}$$

3. *If $\alpha_k$ is given by exact linesearch, then $\alpha_k \ge \bar\alpha_k$ and (5.1) is again satisfied with $\rho = \frac{1}{2}$.*

*Proof.* By the standard descent lemma [5, Proposition 6.1.2] we have

$$f(x_k) - f(x_k + \alpha d_k) \ge \alpha p_k - \alpha^2 \frac{L\|d_k\|^2}{2} \ . \tag{7.2}$$

It is immediate to check

$$\alpha\nabla f(x_k)^\top d_k + \alpha^2 \frac{L\|d_k\|^2}{2} \le 0 \ , \tag{7.3}$$

for every $0 \le \alpha \le \frac{2p_k}{L\|d_k\|^2}$ and

$$\alpha p_k - \alpha^2 \frac{L\|d_k\|^2}{2} \ge \alpha p_k/2 \ge \alpha^2 \frac{L\|d_k\|^2}{2} \tag{7.4}$$

for every $0 \leq \alpha \leq \frac{p_k}{L\|d_k\|^2}$.

1. For every $x \in \operatorname{conv}(x_k, x_{k+1}) \subseteq \left\{x_k + \alpha d_k \mid 0 \leq \alpha \leq \frac{2p_k}{L\|d_k\|^2}\right\}$, we have

$$f(x) = f(x_k + \alpha d_k) \leq f(x_k) + \alpha \nabla f(x_k)^\top d_k + \alpha^2 \frac{L\|d_k\|^2}{2} \leq f(x_k) \ ,$$

where we used (7.2) in the first inequality and (7.3) in the second inequality.

2. We have

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \bar{\alpha}_k p_k/2 \ ,$$

where we have the hypotheses to apply (7.4) since $0 \leq \bar{\alpha}_k \leq \frac{p_k}{L\|d_k\|^2}$. Again by (7.4)

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \bar{\alpha}_k^2 \frac{L\|d_k\|^2}{2} = L\frac{\|x_k - x_{k+1}\|^2}{2} \ .$$

3. If $\alpha_k = \alpha_k^{\max}$ then there is nothing to prove since $\bar{\alpha}_k \leq \alpha_k^{\max}$. Otherwise we have

$$0 = \frac{\partial}{\partial \alpha} f(x_k + \alpha d_k)|_{\alpha=\alpha_k} = d_k^\top (\nabla f(x_k + \alpha_k d_k)) \tag{7.5}$$

and therefore

$$\begin{aligned} -d_k^\top \nabla f(x_k) &= -d_k^\top \nabla f(x_k) + d_k^\top \nabla f(x_k + \alpha_k d_k) = -d_k^\top (\nabla f(x_k) - \nabla f(x_k + \alpha_k d_k)) \\ &\leq L\|d_k\|\|x_k - (x_k + \alpha_k d_k)\| = \alpha_k L\|d_k\|^2 \ , \end{aligned} \tag{7.6}$$

where we used (7.5) in the first equality and the Lipschitz condition in the inequality. From (7.6) it follows

$$\alpha_k \geq \frac{-d_k^\top \nabla f(x_k)}{L\|d_k\|^2} \geq \bar{\alpha}_k$$

and this proves the first claim. As for the second,

$$f(x_k) - f(x_k + \alpha_k d_k) \geq f(x_k) - f(x_k + \bar{\alpha}_k d_k) \geq \frac{\bar{\alpha}_k}{2} p_k \ ,$$

where the first inequality follows from the definition of exact linesearch and the second by point 2 of the lemma. ∎

Corollary 7.1. *Under the hypotheses of Lemma 7.2, assume that $f(x_k)$ is monotonically decreasing and assume that for some subsequence $k(j)$ we have $x_{k(j)+1} = x_{k(j)} + \bar{\alpha}_{k(j)} d_{k(j)}$. Then*

$$\|x_{k(j)} - x_{k(j)+1}\| \to 0 \ .$$

*Proof.* By (7.1) we have

$$f(x_{k(j)}) - f(x_{k(j)+1}) \geq \frac{L}{2}\|x_{k(j)} - x_{k(j)+1}\|^2$$

and the conclusion follows by monotonicity and boundedness. ∎

**7.2. AFW complexity for generic polytopes.** It is well known as anticipated in the introduction that every application of the AFW to a polytope can be seen as an application of the AFW to the probability simplex.

In this section we show the connection between the active set and the face of the polytope exposed by $-\nabla f(y^*)$, where $y^*$ is a stationary point for $f$. We then proceed to show with a couple of examples how the results proved for the probability simplex can be adapted to general polytopes. In particular we

will generalize Theorem 4.1, thus proving that under a convergence assumption the AFW identifies the face exposed by the gradients of some stationary points. An analogous result is already well known for the gradient projection algorithm, and was first proved in [10] building on [9] which used an additional strict complementarity assumption but worked in a more general setting than polytopes, that of convex compact sets with a polyhedral optimal face.
Before stating the generalized theorem we need to introduce additional notation and prove a few properties mostly concerning the generalization of the simplex multiplier function $\lambda$ to polytopes.
Let $P$ be a polytope and $f : P \to \mathbb{R}^n$ be a function with gradient having Lipschitz constant $L$.
To define the AFW algorithm we need a finite set of atoms $\mathcal{A}$ such that $\mathrm{conv}(\mathcal{A}) = P$. As for the probability simplex we can then define for every $a \in \mathcal{A}$ the multiplier function $\lambda_a : P \to \mathbb{R}$ by

$$\lambda_a(y) = \nabla f(y)^\top (a - y) \ .$$

Let finally $A$ be a matrix having as columns the atoms in $\mathcal{A}$, so that $A$ is also a linear transformation mapping $\Delta_{|\mathcal{A}|-1}$ in $P$ with $Ae_i = A^i \in \mathcal{A}$.
In order to apply Theorem 3.1 we need to check that the transformed problem

$$\min\{f(Ax) \mid x \in \Delta_{|\mathcal{A}|-1}\}$$

still has all the necessary properties under the assumptions we made on $f$.
Let $\tilde{f}(x) = f(Ax)$. First, it is easy to see that the gradient of $\tilde{f}$ is still Lipschitz. Also $\lambda$ is invariant under affine transformation, meaning that $\lambda_{A^i}(Ax) = \lambda_i(x)$ for every $i \in [1 : |\mathcal{A}|]$, $x \in \Delta_{|\mathcal{A}|-1}$. Indeed

$$\lambda_{A^i}(Ax) = \nabla f(Ax)^\top (A^i - Ax) = \nabla f(Ax)^\top A(e_i - x) = \nabla \tilde{f}(x)^\top (e_i - x) = \lambda_i(x) \ .$$

Let $Y^*$ be the set of stationary points for $f$ on $P$, so that by invariance of multipliers $\mathcal{X}^* = A^{-1}(Y^*)$ is the set of stationary points for $\tilde{f}$. The invariance of the identification property follows immediately from the invariance of $\lambda$: if the support of the multiplier functions for $f$ restricted to $B$ is $\{A^i\}_{i \in I^c}$, then the support of the multiplier functions for $\tilde{f}$ restricted to $A^{-1}(B)$ is $I^c$.
We now show the connection between the face exposed by $-\nabla f$ and the support of the multiplier function. Let $y^* = Ax^* \in Y^*$ and let

$$P^*(y^*) = \{y \in P \mid \nabla f(y^*)^\top y = \nabla f(y^*)^\top y^*\} = \mathrm{argmax}\{-\nabla f(y^*)^\top y \mid y \in P\} = \mathcal{F}(-\nabla f(y^*))$$

be the face of the polytope $P$ exposed by $-\nabla f(y^*)$. The complementarity conditions for the generalized multiplier function $\lambda$ can be stated very simply in terms of inclusion in $P^*(y^*)$: since $y^* \in P^*(y^*)$ we have $\lambda_a(y^*) = 0$ for every $a \in P^*(y^*)$, $\lambda_a(y^*) > 0$ for every $a \notin P^*(y^*)$. But $P$ is the convex hull of the set of atoms in $\mathcal{A}$ so that the previous relations mean that the face $P^*(y^*)$ is the convex hull of the set of atoms for which $\lambda_a(y^*) = 0$:

$$P^*(y^*) = \mathrm{conv}\{a \in \mathcal{A} \mid \lambda_a(y^*) = 0\}$$

or in other words since $\lambda_{A^i}(y^*) = 0$ if and only if $i \in I(x^*) = \{i \in [1 : n] \mid \lambda_i(x^*) = 0\}$:

$$P^*(y^*) = \mathrm{conv}\{a \in \mathcal{A} \mid a = A^i, \ i \in I(x^*)\} \ . \tag{7.7}$$

A consequence of (7.7) is that given any subset $B$ of $P$ with a constant active set, we necessarily get $P^*(w) = P^*(z)$ for every $w, z \in B$, since $I(w) = I(z)$. For such a subset $B$ we can then define

$$P^*(B) = P^*(y^*) \text{ for any } y^* \in B$$

where the definition does not depend on the specific $y^* \in B$ considered. We can now restate Theorem 4.1 in slightly different terms:

THEOREM 7.1. *Let $\{y_k\}$ be a sequence generated by the AFW on $P$ and let $\{x_k\}$ be the corresponding sequence of weights in $\Delta_{|\mathcal{A}|-1}$ such that $\{y_k\} = \{Ax_k\}$. Assume that the stepsizes satisfy $\alpha_k \geq \bar{\alpha}_k$ (using $\tilde{f}$ instead of $f$ in (4.2)). If there exists a compact subset $B$ of $Y^*$ with the SIP such that $y_k \to B$, then there exists $M$ such that*

$$y_k \in P^*(B) \text{ for every } k \geq M.$$

*Proof.* Follows from Theorem 4.1 and the affine invariance properties discussed above. ∎

A technical point concerning Theorem 7.1 is that in order to compute $\bar{\alpha}_k$ the Lipschitz constant $L$ of $\nabla\tilde{f}$ (defined on the simplex) is necessary. When optimizing on a general polytope, the calculation of an accurate estimate of $L$ for $\tilde{f}$ may be problematic. However, by Lemma 7.2 if the AFW uses exact linesearch, the stepsize $\bar{\alpha}_k$ (and in particular the constant $L$) is not needed because the inequality $\alpha_k \geq \bar{\alpha}_k$ is automatically satisfied.

We now generalize the analysis of the strongly convex case. The technical problem here is that strong convexity, which is used in Corollary 4.1, is not maintained by affine transformations, so that instead we will have to use a weaker error bound condition. As a possible alternative, in [28] linear convergence of the AFW is proved with dependence only on affine invariant parameters, so that any version of Theorem 3.1 and Corollary 4.1 depending on those parameters instead of $u_1, L$ would not need this additional analysis.

Let $P = \{y \in \mathbb{R}^n \mid Cy \leq b\}$, $y^*$ be the unique minimizer of $f$ on $P$ and $u > 0$ be such that

$$f(y) \geq f(y^*) + \frac{u}{2}\|y - y^*\|^2 \ .$$

The function $\tilde{f}$ inherits the error bound condition necessary for Corollary 4.1 from the strong convexity of $f$: for every $x \in \Delta_{|\mathcal{A}|-1}$ by [3], Lemma 2.2 we have

$$\operatorname{dist}(x, \mathcal{X}^*) \leq \theta\|Ax - y^*\|$$

where $\theta$ is the Hoffman constant related to $[C^T, [I; e; -e]^T]^T$. As a consequence if $\tilde{f}^*$ is the minimum of $\tilde{f}$

$$\tilde{f}(x) - \tilde{f}^* = f(Ax) - f(y^*) \geq \frac{u}{2}\|Ax - y^*\|^2 \geq \frac{u}{2\theta^2}\operatorname{dist}(x, \mathcal{X}^*)^2$$

and using that $n\|\cdot\|^2 \geq \|\cdot\|_1^2$ we can finally retrieve an error bound condition with respect to $\|\cdot\|_1$:

$$\tilde{f}(x) - \tilde{f}^* \geq \frac{u}{2n\theta^2}\operatorname{dist}_1(x, \mathcal{X}^*)^2. \tag{7.8}$$

Having proved this error bound condition for $\tilde{f}$ we can now generalize (3.5):

COROLLARY 7.2. *The sequence $\{y_k\}$ generated by the AFW is in $P^*(y^*)$ for*

$$k \geq \max\left(0, \frac{\ln(h_0) - \ln(u_P r_*^2/2)}{\ln(1/q)}\right) + |I^c|$$

*where $q \in (0,1)$, is the constant related to the linear convergence rate $f(y_k) - f(y^*) \leq q^k(f(y_0) - f(y^*))$, $u_P = \frac{u}{2n\theta^2}$, $r_* = \frac{\delta_{\min}}{2L+\delta_{\min}}$ with $\delta_{\min} = \min\{\lambda_a(y^*) \mid \lambda_a(y^*) > 0\}$.*

*Proof.* Let $I = \{i \in [1 : |\mathcal{A}|] \mid \lambda_{A^i}(y^*) = 0\}$, $P^* = P^*(y^*)$. Since $P^* = \operatorname{conv}(\mathcal{A} \cap P^*)$ and by (7.7) $\operatorname{conv}(\mathcal{A} \cap P^*) = \operatorname{conv}\{A^i \mid i \in I\}$ the theorem is equivalent to prove that for every $k$ larger than the bound, we have $y_k \in \operatorname{conv}\{A^i \mid i \in I\}$. Let $\{x_k\}$ be the sequence generate by the AFW on the probability simplex, so that $y_k = Ax_k$. We need to prove that, for every $k$ larger than the bound, we have

$$x_k \in \operatorname{conv}\{e_i \mid i \in I\} \ ,$$

or in other words $(x_k)_i = 0$ for every $i \in I^c$.

Reasoning as in Corollary 4.1 we get that $\operatorname{dist}_1(x_k, \mathcal{X}^*) < r_*$ for every

$$k \geq \frac{\ln(h_0) - \ln(u_P r_*^2/2)}{\ln(1/q)} \ . \tag{7.9}$$

Let $\bar{k}$ be the minimum index such that (7.9) holds. For every $k \geq \bar{k}$ there exists $x^* \in \mathcal{X}^*$ with $\|x_k - x^*\|_1 < r_*$. But $\lambda_i(x) = \lambda_{A^i}(y^*)$ for every $x \in \mathcal{X}^*$ by the invariance of $\lambda$, so that we can apply Theorem 3.1 with fixed point $x^*$ and obtain that if $J_k = \{i \in I^c \mid (x_k)_i > 0\}$ then $J_{k+1} \leq \max(0, J_k - 1)$. The conclusion follows exactly as in Corollary 4.1. □

## REFERENCES


[1] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

[2] Maxim Balashov, Boris Polyak, and Andrey Tremba. Gradient projection and conditional gradient methods for constrained nonconvex minimization. *arXiv preprint arXiv:1906.11580*, 2019.

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

[4] Dimitri P Bertsekas. Projected newton methods for optimization problems with simple constraints. *SIAM J. Control Optim.*, 20(2):221–246, 1982.

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

[6] Ernesto G Birgin and José Mario Martínez. Large-scale active-set box-constrained optimization method with spectral projected gradients. *Comput. Optim. Appl.*, 23(1):101–125, 2002.

[7] 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.

[8] Immanuel M Bomze, Francesco Rinaldi, and Samuel Rota Buló. 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.

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

[10] James V Burke and Jorge J Moré. Exposing constraints. *SIAM Journal on Optimization*, 4(3):573–595, 1994.

[11] Jim Burke. On the identification of active constraints II: The nonconvex case. *SIAM Journal on Numerical Analysis*, 27(4):1081–1102, 1990.

[12] 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.

[13] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[14] Andrea Cristofari, Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. An Active-Set Algorithmic Framework for Non-Convex Optimization Problems over the Simplex. *arXiv e-prints*, page arXiv:1703.07761, March 2017.

[15] Andrea Cristofari, Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. An active-set algorithmic framework for non-convex optimization problems over the simplex. *arXiv preprint arXiv:1703.07761v2*, 2018.

[16] Marianna De Santis, Gianni Di Pillo, and Stefano Lucidi. An active set feasible method for large-scale minimization problems with bound constraints. *Computational Optimization and Applications*, 53(2):395–423, 2012.

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

[18] Paul Grigas, Alfonso Lobos, and Nathan Vermeersch. Stochastic in-face frank-wolfe methods for non-convex optimization and sparse neural network training. *arXiv preprint arXiv:1906.03580*, 2019.

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

[20] William W Hager, Dzung T Phan, and Hongchao Zhang. Gradient-based methods for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):146–165, 2011.

[21] William W Hager and Hongchao Zhang. A new active set algorithm for box constrained optimization. *SIAM J. Optim.*, 17(2):526–557, 2006.

[22] William W Hager and Hongchao Zhang. An active set algorithm for nonlinear optimization with polyhedral constraints. *Science China Mathematics*, 59(8):1525–1542, 2016.

[23] Warren L Hare and Adrian S Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[24] Alfredo N Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational & Applied Mathematics*, 22(1):37–52, 2003.

[25] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[26] Rahul G Krishnan, Simon Lacoste-Julien, and David Sontag. Barrier frank-wolfe for marginal inference. In *Advances in Neural Information Processing Systems*, pages 532–540, 2015.

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

[28] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

[29] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[30] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent go fast: Faster greedy rules,

message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*, 2017.

[31] 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.

[32] Javier Peña and Daniel Rodriguez. Polytope conditioning and linear convergence of the Frank–Wolfe algorithm. *Mathematics of Operations Research*, 44(1):1–18, 2018.

[33] Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? manifold identification of gradient-related proximal methods. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1110–1119, 2019.

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

[35] Stephen J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.