SIAM J. OPTIM.
Vol. 30, No. 3, pp. 2470–2500

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, for nonconvex objectives, a novel $O(1/\sqrt{k})$ convergence rate result and active set identification for different step sizes (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 an 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

DOI. 10.1137/19M1309419

1. Introduction.

Identifying a surface containing a solution (and/or the support of sparse solutions) represents a relevant task in optimization, since it allows one to reduce the dimension of the problem at hand and apply a more sophisticated method in the end (see, e.g., [5, 8, 17, 18, 22, 23, 24]). This is the reason why, in the last few decades, identification properties of optimization methods have been the subject of extensive studies.

The Frank–Wolfe (FW) algorithm, first introduced in [19], is a classic first order optimization method that has recently regained popularity thanks to the way it can easily handle the structured constraints appearing in many real-world applications. This method and its variants have indeed been applied in the context of, e.g., submodular optimization problems [1], variational inference problems [29], and sparse neural network training [20]. 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., [11]). 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

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

*Received by the editors December 27, 2019; accepted for publication (in revised form) June 4, 2020; published electronically September 16, 2020.
https://doi.org/10.1137/19M1309419
†ISOR, University of Vienna, Vienna, 1090, Austria (immanuel.bomze@univie.ac.at).
‡Dipartimento di Matematica "Tullio Levi-Civita," Università di Padova, Padova, 35121, Italy (rinaldi@math.unipd.it, damiano.zeffiro@math.unipd.it).

is the probability simplex. When the algorithm converges to a stationary point x^* we say that it identifies the active set if it correctly determines all of those constraints whose multiplier is positive at x^* (see (2.1)). The active set complexity is then defined as the number of iterations after which every sequence generated by the algorithm identifies this subset of constraints. In this paper, we extend the active set complexity definition to include sequences convergent to certain subsets of stationary points.

1.1. Contributions. It is a classic result that on polytopes and under strict complementarity conditions the AFW with exact line search identifies the face containing the minimum in finite time for strongly convex objectives [21]. More general active set identification properties for FW variants have recently been analyzed in [11], 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 (possibly neither concave nor convex) quadratic functions. The main contributions of this article with respect to [11] are twofold:

- 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 a constraint whose multiplier is positive (if there are any remaining to be identified still). 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 [11]. Instead, we consider general nonconvex objectives with Lipschitz gradient. More specifically, we prove active set identification under different conditions on the step size and some additional hypotheses on the support of stationary points.

In order to prove our results, we consider step sizes dependent on the Lipschitz constant of the gradient (see, e.g., [2, 26] and references therein). By exploiting the affine invariance property of the AFW (see, e.g., [27]), we also extend some of the results to generic polytopes. In our analysis we 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.

This paper is organized as follows: after presenting the AFW method and the setting in section 2, we study the local behavior 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, section 6, we draw some conclusions. To improve readability, some technical details are deferred to the appendices.

1.2. Related work. In [13] the authors proved that the projected gradient method and other converging sequential quadratic programming methods identify quasi-polyhedral faces under some nondegeneracy conditions. In [14] 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 use in studying the AFW on polytopes. In [39] the results of [13] 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 [13] to nonconvex sets was proved in [12], while an extension to nonsmooth objectives was first proved in [25]. Active set identification results have also been proved for a variety of projected gradient, proximal gradient, and stochastic gradient related methods (see, for instance, [37] 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 methods were analyzed in [35, 34] 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 [37].

As for FW-like methods, in addition to the results in [21, 11] discussed earlier, identification results have been proved in [16] for fully corrective variants on the probability simplex. However, since fully corrective variants require computing the minimum of the objective on a given face at each iteration, they are not suited for nonconvex problems.

2. Preliminaries. In this article, $f : \Delta_{n-1} \to \mathbb{R}$ is a function with a gradient having Lipschitz constant L. The constant L is also used as a Lipschitz constant for ∇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 $x, y \in \Delta_{n-1}$. We denote by $\mathcal{X}^*$ the set of points satisfying first order optimality conditions for the minimization of f on Δ_{n-1}; that is, $\nabla f(x)^\top d \geq 0$ for every d feasible direction at x. We call $\mathcal{X}^*$ the set of stationary points (see, e.g., [6]).

For $x \in \mathbb{R}^n$ and $X \subset \mathbb{R}^n$, the function $\operatorname{dist}(x, X)$ is the standard point-set distance and for $A \subset \mathbb{R}^n$ the function $\operatorname{dist}(A, X)$ is the infimum of the distance between points in the following set:

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

We define dist_1 in the same way but with respect to $\|\cdot\|_1$. We denote with

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

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. For a sequence $\{a^{(k)}\}_{k\in\mathbb{N}_0}$ we drop the subscript and write simply $\{a^{(k)}\}$ (unless, of course, the sequence is defined on some other index set). We use the notation $a^{(k)} \to A$ for the convergence of $\{a^{(k)}\}$ to the set A as equivalent to $\operatorname{dist}(a^{(k)}, A) \to 0$.

We now introduce the multiplier functions, which were recently used in [17] 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 $x_i \geq 0$.

We define *the extended support* in $x \in \mathcal{X}^*$ as

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

and with

$$I^c(x) = \{1, \ldots, .n\} \setminus I(x) \tag{2.1}$$

the set of constraints whose multiplier is positive in x, where by optimality conditions we have $\lambda_i(x) \geq 0$ for every $i \in [1:n]$. Therefore,

$$\lambda_i(x) > 0 \;\forall\; i \in I^c(x)\,.$$

FW variants require a linear minimization oracle (LMO) for the feasible set (the probability simplex in our case):

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

Keeping in mind that

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

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

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

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

Algorithm 2.1 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 step size if stationarity conditions are not satisfied. It is well known [15, 38] that the method exhibits a zigzagging behavior 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.

Algorithm 2.1 FW method on the probability simplex.

1: **Initialize** $x^{(0)} \in \Delta_{n-1}$, $k := 0$
2: Set $s_k := e_{\hat{\imath}}$, with $\hat{\imath} \in \operatorname{argmin}_i \nabla f(x^{(k)})_i$ and $d^{(k)}_{\mathcal{FW}} := s_k - x^{(k)}$
3: **if** $x^{(k)}$ is stationary **then**
4: STOP
5: **end if**
6: Choose the step size $\alpha_k \in (0, 1]$ with a suitable criterion
7: Update: $x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}_{\mathcal{FW}}$
8: Set $k := k + 1$. Go to step 2.

Algorithm 2.2 AFW on the probability simplex.

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

Both of these issues are solved by the away-step variant of the FW method, reported in Algorithm 2.2. The AFW at every iteration chooses between the classic FW direction and the away-step direction $d^{(k)}_{\mathcal{A}}$ calculated in Step 6. 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 approximation of f given by $y \to f(x^{(k)}) + \nabla f(x^{(k)})^\top (y - x^{(k)})$. The step size upper bound $\alpha^{\max}_k$ in Step 12 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 by one. In the latter case $\alpha_k = \alpha^{\max}_k$ and we get rid of the component whose index is associated to the away direction. On the other hand, when the algorithm performs an FW step, only the vertex given by the LMO can be added to the support of the current iterate. These two properties are fundamental for the active set identification of the AFW.

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.

Let $\{x^{(k)}\}$ be the sequence of points generated by the AFW, and let x^* be a fixed point in $\mathcal{X}^*$. We write for simplicity I and I^c instead of $I(x^*)$ and $I^c(x^*)$, respectively, in the rest of this section, since x^* does not change. Note that by complementary slackness we have $x^*_j = 0$ for all $j \in I^c$.

The first result of this section is a technical lemma that allows us 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 $\|x^{(k)} - x^*\|_1 \le h$, let*

$$O_k = \{i \in I^c \mid x_i^{(k)} = 0\},$$

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

$$|\lambda_i(x^*) - \lambda_i(x^{(k)})| \le h\Big(L + \frac{\delta_k}{2}\Big)\,. \tag{3.1}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We can finally write

$$
|\lambda_i(x^{(k)}) - \lambda_i(x^*)| \le h\Big(L + \frac{\delta_k}{2}\Big)\,,
$$

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. 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 Δ_{n-1} minus the constant $\nabla f(x^{(k)})^\top x^{(k)}$, 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. *Using the notation introduced in Algorithm* 2.2*, we have the following:*

(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* $d^{(k)} = d_{\mathcal{A}}^{(k)} = x^{(k)} - e_{\hat{\imath}}$ *for some* $\hat{\imath} \in \operatorname{argmax}\{\lambda_i(x^{(k)}) \mid i \in S_k\}$.

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

Proof. (a) By the definition of the away direction $d_{\mathcal{A}}^{(k)}$ it follows that

$$
d_{\mathcal{A}}^{(k)} \in \operatorname{argmax}\{-\nabla f(x^{(k)})^\top d \mid d = x^{(k)} - e_i, i \in S_k\}\,,
$$

which implies

$$
\begin{aligned}
d_{\mathcal{A}}^{(k)} &= x^{(k)} - e_{\hat{\imath}} \quad \text{for some } \hat{\imath} \in \operatorname{argmax}\{-\nabla f(x^{(k)})^\top (x^{(k)} - e_i) \mid i \in S_k\} \\
&= \operatorname{argmax}\{\lambda_i(x^{(k)}) \mid i \in S_k\}\,.
\end{aligned} \tag{3.5}
$$

As a consequence of (3.5)

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

where the second equality follows from $\lambda_i(x^{(k)}) = -\nabla f(x^{(k)})^\top d$ with $d = x^{(k)} - e_i$.

Analogously,

$$-\nabla f(x^{(k)})^\top d_{\mathcal{FW}}^{(k)} = \max\{-\nabla f(x^{(k)})^\top d \mid d = e_i - x^{(k)}, i \in \{1,\dots,n\}\} \\ = \max\{-\lambda_i(x^{(k)}) \mid i \in \{1,\dots,n\}\}\,. \tag{3.7}$$

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

$$-\nabla f(x^{(k)})^\top d_{\mathcal{FW}}^{(k)} = \max\{-\lambda_i(x^{(k)}) \mid i \in \{1,,\dots,n\}\} \\ < \max\{\lambda_i(x^{(k)}) \mid i \in S_k\} = -\nabla f(x^{(k)})^\top d_{\mathcal{A}}^{(k)}\,,$$

where we used (3.6) and (3.7) for the first and second equality, respectively, and the inequality is true by hypothesis.

(b) By considering the fact that $x_i^{(k)} = 0$, 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 FW direction at step k as well. This guarantees that $x_i^{(k+1)} = 0$. ∎

For $x \in \mathcal{X}^*$ such that $I^c(x) \neq \emptyset$, we define $\delta_{\min}(x)$ as

$$\delta_{\min}(x) = \min_{i \in I^c(x)} \lambda_i(x)$$

and the active set radius $r_*(x)$ as

$$r_*(x) = \begin{cases} \frac{\delta_{\min}(x)}{\delta_{\min}(x)+2L} & \text{if } I^c(x) \neq \emptyset, \\ +\infty & \text{if } I^c(x) = \emptyset\,. \end{cases}$$

In the rest of this section, we write r_* and $\delta_{\min}$ instead of $r_*(x^*)$ and $\delta_{\min}(x^*)$. Having introduced these constants, we can now state the AFW local identification theorem.

THEOREM 3.3. *Assume that for every k such that $d^{(k)} = d_{\mathcal{A}}^{(k)}$ the step size α_k is either maximal with respect to the boundary conditions (that is, $\alpha_k = \alpha_k^{\max}$) or $\alpha_k \geq \frac{-\nabla f(x^{(k)})^\top d^{(k)}}{L\|d^{(k)}\|^2}$. If $\|x^{(k)} - x^*\|_1 < r_*$ then*

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

The latter relation also holds in the case $I^c = \emptyset$.

In the proof, we split $[1:n]$ into 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. 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 extended support of the stationary point, then we show that the AFW chooses a direction contained in the extended support, so that also $J_{k+1} = \emptyset$. If $J_k \neq \emptyset$, we 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.

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 $I^c \neq \emptyset$. Recall that $\lambda_i(x^*) > 0$ for every $i \in I^c$, so that necessarily $\delta_{\min} > 0$.

For every $i \in [1:n]$, by Lemma 3.1

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

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 $x_i^{(k)} = 0$ 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 $x_i^{(k+1)} = 0$ for every $i \in I^c$. Hence $J_{k+1} = \emptyset = J_k$ and Theorem 3.3 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 δ_k, it follows that

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

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 $\delta_{\min} \mapsto -\frac{\delta_{\min}}{2L + \delta_{\min}}$ 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 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

$$\lambda_j(x^{(k)}) > \delta_{\min} - \frac{\delta_{\min}(L + \frac{\delta_k}{2})}{2L + \delta_{\min}} > -\frac{\delta_k}{2}\,.$$

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 $d^{(k)} = d_{\mathcal{A}}^{(k)} = x^{(k)} - e_{\hat{\jmath}}$ 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 $x_{\hat{\jmath}}^{(k)} = 0$ 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 step size 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 A.1

(3.17)

$$\|d^{(k)}\| = \|e_{\hat{\jmath}} - x^{(k)}\| \leq \sqrt{2}(e_{\hat{\jmath}} - x^{(k)})_{\hat{\jmath}} = -\sqrt{2}d_{\hat{\jmath}}^{(k)} \Rightarrow \frac{d_{\hat{\jmath}}^{(k)}}{\|d^{(k)}\|^2} \leq \frac{d_{\hat{\jmath}}^{(k)}}{\sqrt{2}\|d^{(k)}\|} \leq -1/2,$$

$$x_{\hat{\jmath}}^{(k)} = (x^{(k)} - x^*)_{\hat{\jmath}} \leq \frac{\|x^{(k)} - x^*\|_1}{2} < \frac{r_*}{2} = \frac{\delta_{\min}}{4L + 2\delta_{\min}}\,.$$

Finally, combining (3.17) with (3.16),

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

where we used (3.16) to bound α_k in the first inequality, (3.17) to bound $x_{\hat{\jmath}}^{(k)}$ and $\frac{d_{\hat{\jmath}}^{(k)}}{\|d^{(k)}\|^2}$. Hence $x_{\hat{\jmath}}^{(k+1)} < 0$, a 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 $I^c(x^*)$ related to those constraints whose multipliers are positive in x^*. In general, $I^c(x^*) \subset \operatorname{supp}(x^*)^c$ by complementarity conditions, with

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

The face $\mathcal{F}$ of Δ_{n-1} defined by the constraints with indices in $I^c(x^*)$ has a nice geometrical interpretation: it is the face of Δ_{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 $I^c_A \subset [1:n]$ such that*

$$I^c(x) = I^c_A \quad \forall\, x \in A\,.$$

In other words, A has the SIP if and only if $I^c(x)$ or, equivalently, the extended support $I(x)$ is constant for x varying in A. The geometrical interpretation of Definition 4.1 is the following: for every point x in the subset A, the negative gradient $-\nabla f(x)$ exposes the same face. This is trivially true if A is a singleton so that the notion of subset with the SIP generalizes the one of stationary point. From the geometrical interpretation it is clear that A has the SIP also if it is contained in the relative interior of a face $\mathcal{F}$ of Δ_{n-1} and strict complementarity conditions hold for every point in A. In this case the negative gradient of the points in A always exposes $\mathcal{F}$. As a pathological example, for $f \equiv 0$ all the subsets of Δ_{n-1} have the SIP because every $x \in \Delta_{n-1}$ is stationary with $I^c(x) = \emptyset$.

For a set A with the SIP we define

$$r_*(A) = \min_{x \in A} r_*(x)\,.$$

Thanks to the SIP, r_* is continuous on A and we always have $r_*(A) > 0$. 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 say that this algorithm solves the active set problem in M steps if $x_i^{(k)} = 0$ for every $i \in I^c_A$, $k \geq M$. If, given a set of conditions on $(A, f, x^{(0)})$, M is the minimum number which has this property for every sequence generated by the algorithm, then we say that the active set complexity of the algorithm is M, under the given conditions.*

We can now apply Theorem 3.3 to show that once a sequence is definitely close enough to a set A enjoying the SIP, the AFW identifies the active set in at most $|I^c_A|$ steps. We first need to define a quantity that we use as a lower bound on the step sizes:

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

THEOREM 4.3. *Let $\{x^{(k)}\}$ be a sequence generated by the AFW, with step size $\alpha_k \geq \bar{\alpha}_k$. Let $\mathcal{X}^*$ be the set of stationary points of a function $f : \Delta_{n-1} \to \mathbb{R}$ with ∇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*

$$x_i^{(k)} = 0 \quad \text{for every } k \geq M \text{ and all } i \in I^c_A\,.$$

We refer the reader to Remark 4.4 for some examples of step size strategies satisfying (4.2).

Proof. Let $J_k = \{i \in I_A^c \mid x_i^{(k)} > 0\}$ and choose $\bar{k}$ such that $\text{dist}_1(x^{(k)}, A) < r_*(A)$ for every $k \geq \bar{k}$. Then for every $k \geq \bar{k}$ there exists $y^* \in A$ with $\|x^{(k)} - y^*\|_1 < r_*(A) \leq r_*(y^*)$. Since by hypothesis for every $y^* \in A$ the support of the multiplier function is I_A^c applying Theorem 3.3 with y^* as a fixed point, we 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 $|J_{\bar{k}+|I_A^c|}| = 0$. ∎

Remark 4.4. In Appendix B we prove that (4.2) is always a lower bound on the step size obtained by the exact line search. We also prove that

$$\alpha_k \geq \min\left(\alpha_k^{\max}, c\frac{p_k}{L\|d^{(k)}\|^2}\right) \text{ for some } c > 0$$

for the Armijo line search, and if we impose the weak Wolfe conditions, setting $\alpha_k = \alpha_k^{\max}$ whenever those conditions cannot be satisfied. When $c \geq 1$, then (4.2) is, of course, a lower bound for the step size α_k, and when $c < 1$ we can still recover (4.2) by considering $\tilde{L} = \frac{L}{c}$ instead of L as Lipschitz constant.

The proof of Theorem 4.3 also gives a relatively simple upper bound for the complexity of the active set problem.

PROPOSITION 4.5. *Under the assumptions of Theorem* 4.3*, the active set complexity is at most*

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

We now report an explicit bound for the strongly convex case, and will analyze in depth the nonconvex case later in section 5. If f is u-strongly convex, then $f(x^*)$ is the global minimum of f over Δ^{n-1} if x^* is the (unique) stationary point; further, it is easy to see that the following inequality holds for every x on Δ_{n-1}:

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

with $u_1 = u/n$.

COROLLARY 4.6. *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* $h_k = f(x^{(k)}) - f(x^*)$*, be the convergence rate related to the AFW (see* [31, Theorem 8]*). Then the active set complexity is*

$$\max\left(0, \left\lceil \frac{\ln(h_0) - \ln(u_1 r_*(x^*)^2/2)}{\ln(1/\mathrm{q})} \right\rceil\right) + |I^c|\,. \tag{4.5}$$

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

$$h_k \leq \frac{u_1}{2} r_*(x^*)^2 \tag{4.6}$$

is at most

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

We claim that if condition (4.6) holds, then it takes at most $|I^c|$ steps for the sequence to be definitely in the active set. Indeed, if $q^k h_0 \leq \frac{u_1}{2} r_*(x^*)^2$, then necessarily $x^{(k)} \in B_1(x^*, r_*(x^*))$ by (4.3), and by monotonicity of the bound (4.4) we then have $x^{(k+h)} \in B_1(x^*, r_*(x^*))$ for every $h \geq 0$. Once the sequence is definitely in $B_1(x^*, r_*(x^*))$ 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 4.7. In Corollary 4.6, if we assume the linear rate (4.4) (which may not hold in the nonconvex case), then the strong convexity of f can be replaced by the condition (4.3).

An extension of Corollary 4.6 to generic polytopes, requiring additional theoretical results, is presented in Appendix C.

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 is 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 if and only if x is a stationary point. The reason why this function is called an 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., [27]) and is computed for free at each iteration of an FW-like algorithm. In [30], 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 [32] 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 [36], we now study the AFW convergence rate in the nonconvex case with respect to the sequence $\{g^*_k\}$.

5.1. Global convergence. We start investigating the minimal FW gap, giving estimates of rates of convergence. In the next theorem and in the subsequent corollary, Corollary 5.2, we assume that the AFW starts from a vertex of the probability simplex. Thanks to the affine invariance properties of the AFW this is not a restrictive assumption. For a generic starting point one can indeed apply the same theorem to the AFW starting from e_{n+1} for $\tilde{f} : \Delta_n \to \mathbb{R}$ satisfying

$$\tilde{f}(y) = f(y_1 e_1 + \cdots + y_n e_n + y_{n+1} x^{(0)})\,, \tag{5.1}$$

where $x^{(0)} \in \Delta_{n-1}$ is the desired starting point (see also Corollary 5.3). Formally, this leads to the computation of a sequence $\{y^{(k)}\}$ on Δ_n which can be mapped to a sequence $\{x^{(k)}\}$ on Δ_{n-1} by the affine transformation

$$p(y) = y_1 e_1 + \cdots + y_n e_n + y_{n+1} x^{(0)} . \tag{5.2}$$

In Appendix C, we discuss the invariance of the AFW under affine transformations in more detail.

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 Δ_{n-1}, with $x^{(0)}$ a vertex of Δ_{n-1}. Assume that the step size α_k is equal to or greater than $\bar{\alpha}_k$ (as defined in (4.2)), and that*

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \geq \rho \bar{\alpha}_k \left(-\nabla f(x^{(k)})^\top d^{(k)} \right) \tag{5.3}$$

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

$$g_T^* \leq \max \left(\sqrt{\frac{4L(f(x^{(0)}) - f^*)}{\rho T}}, \frac{4(f(x^{(0)}) - f^*)}{T} \right) . \tag{5.4}$$

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 (5.3) 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,

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\rho}{L\|d^{(k)}\|^2} (r_k^\top d^{(k)})^2 \geq \frac{\rho}{L\|d^{(k)}\|^2} g_k^2 \geq \frac{\rho g_k^2}{2L} , \tag{5.5}$$

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 [7, 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

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

hence we can write

$$f(x^{(k)}) - f(x^{(k+1)}) \geq -\nabla f(x^{(k)})^\top d^{(k)} - \frac{L}{2}\|d^{(k)}\|^2 \geq -\frac{\nabla f(x^{(k)})^\top d^{(k)}}{2} \geq \frac{1}{2} g_k . \tag{5.6}$$

Reasoning as in Case 1 we also have $|S_{k+1}| \leq |S_k| + 1$.

Case 3. $\alpha_k = \bar{\alpha}_k = \alpha_k^{\max}$, $d^{(k)} = d_{\mathcal{A}}^{(k)}$. Then $d^{(k)} = x^{(k)} - e_i$ for $i \in S_k$ and

$$x_j^{(k+1)} = (1+\alpha_k)x_j^{(k)} - \alpha_k (e_i)_j \,,$$

with $\alpha_k = \alpha_k^{\max} = \frac{x_i^{(k)}}{1-x_i^{(k)}}$. Therefore, $x_j^{(k+1)} = 0$ for $j \in [1:n] \setminus S_k \cup \{i\}$ and $x_j^{(k+1)} \neq 0$ for $j \in S_k \setminus \{i\}$. In particular, $|S_{k+1}| = |S_k| - 1$.

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

$$|S_T| - |S_0| \leq n_1(T) + n_2(T) - n_3(T) \tag{5.7}$$

for every $T \in \mathbb{N}$.

Since $n_3(T) = T - n_1(T) - n_2(T)$, from (5.7) we get

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T}{2} \,, \tag{5.8}$$

where we used $|S_0| = 1 \leq |S_T|$. Now let C_i^T be the set of iteration counters up to $T-1$ corresponding to Case 1 steps for $i \in \{1,2,3\}$, which satisfies $|C_i^T| = n_i(T)$. We have by summing (5.5) and (5.6) for the indices in C_1^T and C_2^T, respectively,

$$\sum_{k \in C_1^T} f(x^{(k)}) - f(x^{(k+1)}) + \sum_{k \in C_2^T} f(x^{(k+1)}) - f(x^{(k)}) \geq \sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \,. \tag{5.9}$$

We now lower bound the right-hand side of (5.9) in terms of g_T^* as follows:

$$\begin{aligned} &\sum_{k \in C_1^T} \frac{\rho g_k^2}{2L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \geq |C_1^T| \min_{k \in C_1^T} \frac{\rho g_k^2}{2L} + |C_2^T| \min_{k \in C_2^T} \frac{g_k}{2} \\ &\geq (|C_1^T| + |C_2^T|) \min\left(\frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) = [n_1(T) + n_2(T)] \min\left(\rho \frac{(g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) \\ &\geq \frac{T}{2} \min\left(\frac{\rho (g_T^*)^2}{2L}, \frac{g_T^*}{2} \right) . \end{aligned} \tag{5.10}$$

Since the left-hand side of (5.9) 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

$$g_T^* \leq \frac{4(f(x^{(0)}) - f^*)}{T}, \tag{5.11}$$

and otherwise,

$$g_T^* \leq \sqrt{\frac{4L(f(x^{(0)}) - f^*)}{\rho T}} \,. \tag{5.12}$$

The claim follows by taking the max in the system formed by (5.11) and (5.12). ∎

In Appendix B, we prove that condition (5.3) is satisfied by exact line search and Armijo line search as well. We also prove that it is satisfied if we impose the weak Wolfe conditions and take $\alpha_k^{\max}$ whenever the conditions are incompatible with the constraint $\alpha_k \leq \alpha_k^{\max}$.

When the step sizes coincide with the lower bounds $\bar{\alpha}_k$ or are obtained using exact line search, we have the following corollary.

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1*, if* $\alpha_k = \bar{\alpha}_k$ *or if* α_k *is selected by exact line search, then for every* $T \in \mathbb{N}$*,*

$$g_T^* \leq \max\left(\sqrt{\frac{8L(f(x^{(0)}) - f^*)}{T}}, \frac{4(f(x^{(0)}) - f^*)}{T}\right). \tag{5.13}$$

Proof. By points 2 and 3 of Lemma B.1, relation (5.3) is satisfied with $\rho = \frac{1}{2}$ for both $\alpha_k = \bar{\alpha}_k$ and α_k given by exact line search, and we also have $\alpha_k \geq \bar{\alpha}_k$ in both cases. The conclusion follows directly from Theorem 5.1. ∎

Applying the trick of adding the starting point as a vertex allows us to drop the assumptions of starting from a vertex in Theorem 5.1.

COROLLARY 5.3. *Let* $x^{(0)} \in \Delta_{n-1}$*, and let* $\{y^{(k)}\}$ *be a sequence generated by the AFW applied to the objective function* $\tilde{f}$ *defined in* (5.1) *with* $y^{(0)} = e_{n+1}$*. Let* $\{x^{(k)}\} = \{p(y^{(k)})\}$*, for the transformation* p *defined in* (5.2)*. Then under the assumptions of Theorem* 5.1 *on* α_k *and* f*, the bound* (5.4) *and Corollary* 5.2 *still hold.*

Proof. The multipliers are invariant by affine transformation (see Appendix C for further details), and since the FW gap depends on the multipliers, it is also invariant under affine transformation. Also adding the multiplier related to $x^{(0)}$ does not change the FW gap, which is always realized in one of the vertices of the original simplex since it is the maximum of a linear function plus a constant. Therefore, the FW gap is invariant with respect to the transformation p, so that the same arguments used for Theorem 5.1 and Corollary 5.2 can still be applied to $\{x^{(k)}\} = \{p(y^{(k)})\}$. ∎

Since adding a vertex alters the active set identification properties of the problem (e.g., the active set radius), we cannot apply the above results directly in the rest of this article. Instead we use some key intermediate results presented in the proof of Theorem 5.1.

5.2. A general active set identification result. In this section we give a general active set identification result in the nonconvex setting. When the step sizes do not coincide with the lower bound (4.2) we need strict complementarity in this context. 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

$$\operatorname{supp}(x) = [1:n] \setminus I^c(x) = [1:n] \setminus I_A^c \tag{5.14}$$

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

For the result we need an observation on connectedness which seems to be folklore in an optimization context. This property is needed, e.g., for the proof of [32, Theorem 4.1.2] and similar results are discussed in [3]. However, we are not aware of an explicit proof for this property, so for the readers' convenience we provide a short argument.

LEMMA 5.4. *Let* $\{x^{(k)}\}$ *be a bounded sequence in* $\mathbb{R}^n$ *such that* $\|x^{(k)} - x^{(k+1)}\| \to 0$*. Then the set of limit points of* $\{x^{(k)}\}$ *is connected.*

Proof. Assume by contradiction that there are two open sets U_1 and U_2 separating the limit points of $\{x^{(k)}\}$. Then there must exist an infinite number of points from $\{x^{(k)}\}$ both in U_1 and U_2, and in particular, a subsequence $\{x^{(k(j))}\}$ of $\{x^{(k)}\}$ such that $x^{(k(j))} \in U_1$ and $x^{(k(j)+1)} \in U_1^c$ for every $j \in \mathbb{N}_0$. By the condition $\|x^{(k(j))} - x^{(k(j)+1)}\| \to 0$ we obtain

$$\operatorname{dist}(x^{(k(j))}, U_1^c) \to 0\,. \tag{5.15}$$

Since $\{x^{(k(j))}\}$ is bounded by hypothesis it has a nonempty set of limit points. But every limit point of $\{x^{(k(j))}\}$ must be necessarily in U_1^c by (5.15) and also in the closure of U_1 (because $\{x^{(k(j))}\} \subset U_1$) and therefore not in U_2, a contradiction. ∎

We proceed with the announced result.

THEOREM 5.5. *Let $\{x^{(k)}\}$ be the sequence generated by the AFW method with step sizes satisfying $\alpha_k \geq \bar{\alpha}_k$ and* (5.3)*, where $\bar{\alpha}_k$ is given by* (4.2)*. Let $\mathcal{X}^*$ be the subset of stationary points of f. We have the following:*

(a) *$x^{(k)} \to \mathcal{X}^*$ as $k \to \infty$ without any further assumptions.*

(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_A^c in finite time.*

Assume now that $\mathcal{X}^* = \bigcup_{i=1}^C A_i$ with A_i compact for each $i \in [1:C]$, 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 $I_{A_l}^c$ in finite time.*

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

$$x^{(k(j))} \to g^{-1}(0) = \mathcal{X}^*\,, \tag{5.16}$$

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 B.2 to obtain

$$\|x^{(k'(j))} - x^{(k'(j)+1)}\| \to 0\,. \tag{5.17}$$

Combining (5.16), (5.17), 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 B.1, if $\alpha_k = \bar{\alpha}_k$, then $\|x^{(k+1)} - x^{(k)}\| \to 0$. Now Lemma 5.4 together with point (a) ensures that the set of limit points must be contained in a connected component A of $\mathcal{X}^*$. By Theorem 4.3 it follows that if A has the SIP, then $\{x^{(k)}\}$ identifies I_A^c in finite time.

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

$$r_i < r_*(A_i)$$

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

$$\|x - x^*\|_1 < r_*(x^*)\,. \tag{5.18}$$

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

$$\operatorname{supp}(y) \supseteq \operatorname{supp}(A_i) \ \forall y \in U_i\,, \tag{5.19}$$

where these conditions can always be 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 $\{A_i\}_{i=1}^C$. In particular, let A_l minimize $|\operatorname{supp}(A_l)|$ 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 3.3, where by (5.18) the hypotheses of Theorem 3.3 are satisfied for $k = t(j)$ and some $x^* \in A_l$.

Case 1. $x_h^{(t(j))} = 0$ for every $h \in I^c_{A_l}$. In the notation of Theorem 3.3 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 $x_h^{(t(j)+1)} = x_h^{(t(j))} = 0$ for every $h \in I^c_{A_l}$ by Lemma 3.2, so that we can write

$$\operatorname{supp}(A_m) \subseteq \operatorname{supp}(x^{(t(j)+1)}) \subseteq [1:n] \setminus I^c_{A_l} = \operatorname{supp}(A_l)\,, \tag{5.20}$$

where the first inclusion is justified by (5.19) for $i = m$ and the second by strict complementarity (see also (5.14) and the related discussion). But since by hypothesis $\operatorname{supp}(A_m) \neq \operatorname{supp}(A_l)$ the inclusion (5.20) 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 3.3 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,

$$\tilde d_{\bar h} = -\sum_{h\in[1:n]\setminus\bar h} \tilde d_h = -\sum_{h\in[1:n]\setminus\bar h} |\tilde d_h|\,. \tag{5.21}$$

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

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

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 (5.21).

Reasoning by contradiction we have proved that all of the limit points of $\{x^{(k)}\}$ are in A_l for some $l \in [1,\dots,C]$. The conclusion follows immediately from Theorem 4.3. ∎

5.3. Quantitative version of active set identification. Let $q : \mathbb{R}_{>0} \to \mathbb{N}_0$ be such that $f(x^{(k)}) - f(x^{(k+1)}) \le \varepsilon$ for every $k \ge q(\varepsilon)$. In this section, we give global active set complexity bounds for nonconvex objectives as a function of q, which measures how long it takes for $\gamma_k = f(x^{(k)}) - f(x^{(k+1)})$ to fall definitely under a threshold value. We assume that the gap function $g(x)$ satisfies the Hölderian error bound condition

$$g(x) \ge \theta \, \mathrm{dist}_1(x, \mathcal{X}^*)^p \tag{5.22}$$

for some $\theta, p > 0$. This condition is satisfied, e.g., if $f(x)$ (and therefore, $\nabla f(x)$) is a semialgebraic function. In this case then $g(x)$ is also semialgebraic because it is obtained by sums, products, and maxima of semialgebraic functions, and (5.22) holds by Łojasiewicz' inequality (Corollary 2.6.7 in [9]; see also [10] and references therein) applied to g and $\mathrm{dist}_1(x, \mathcal{X}^*)$.

In the convex case, condition (5.22) on the FW gap $g(x)$ is *weaker* than the more common Hölderian error bound condition on the objective; see [10, 28, 40]. This follows trivially from the fact that the FW gap $g(x)$ is always larger than the objective gap $f(x) - f^*$ for convex f. The Hölderian error bound assumption on the gap allows us to give more explicit active set complexity bounds.

THEOREM 5.6. *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 $0 < d \stackrel{\text{def}}{=} \min_{\{i,j\}\subset[1:C]} \mathrm{dist}_1(A_i, A_j)$. Let $\bar{r}_*$ be the minimum active set radius of the sets $\{A_i\}_{i=1}^C$. Assume that $g(x)$ satisfies (5.22). Assume that the step sizes satisfy $\alpha_k = \bar{\alpha}_k$, with $\bar{\alpha}_k$ given by (4.2). Then the active set complexity is at most $q(\bar{\varepsilon}) + n - 1$ for $\bar{\varepsilon}$ satisfying the following conditions:*

$$\bar{\varepsilon} < L\,, \quad \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} < \bar{r}_*, \quad \textit{and} \quad 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} + 2n\sqrt{\frac{2\bar{\varepsilon}}{L}} \le d\,. \tag{5.23}$$

The proof is substantially a quantitative version of the argument used to prove point (b) of Theorem 5.5.

Proof. Fix $k \ge q(\bar{\varepsilon})$, so that

$$f(x^{(k)}) - f(x^{(k+1)}) \le \bar{\varepsilon}\,. \tag{5.24}$$

We refer to Case 1 steps for $i \in [1:3]$ following the definitions in Theorem 5.1. If the step k is a Case 1 step, then by (5.5) with $\rho = 1/2$ we have

$$f(x^{(k)}) - f(x^{(k+1)}) \ge \frac{g(x^{(k)})^2}{4L},$$

and this together with (5.24) implies

$$2\sqrt{L\bar{\varepsilon}} \ge 2\sqrt{L(f(x^{(k)}) - f(x^{(k+1)}))} \ge g(x^{(k)})\,.$$

Analogously, if the step k is a Case 2 step, then by (5.6) we have

$$f(x^{(k)}) - f(x^{(k+1)}) \ge \frac{g(x^{(k)})}{2}$$

so that $2\bar{\varepsilon} \geq g(x^{(k)})$. By the leftmost condition in (5.23) 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

$$g(x^{(k)}) \leq 2\sqrt{L\bar{\varepsilon}}\,. \tag{5.25}$$

By inverting relation (B.1), we also have

$$\|x^{(k)} - x^{(k+1)}\| \leq \sqrt{\frac{2(f(x^{(k)}) - f(x^{(k+1)}))}{L}} \leq \sqrt{\frac{2\bar{\varepsilon}}{L}}\,. \tag{5.26}$$

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 (5.25),

$$\operatorname{dist}_1(x^{(\bar{k})}, \mathcal{X}^*) \leq \left(\frac{g(x^{(\bar{k})})}{\theta}\right)^{\frac{1}{p}} \leq \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} < \bar{r}_*\,, \tag{5.27}$$

where we used (5.25) in the second inequality and the second condition of (5.23) 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 now claim that $I^c_{A_l}$ is already identified at the step $\bar{k}$.

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 (5.27). 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,

$$\begin{aligned}\|x^{(s(k'))} - x^{(s(k'+1))}\|_1 &\leq \sum_{i=s(k')}^{s(k'+1)-1} \|x^{(i+1)} - x^{(i)}\|_1 \leq 2 \sum_{i=s(k')}^{s(k'+1)-1} \|x^{(i+1)} - x^{(i)}\| \\ &\leq 2[s(k'+1) - s(k')]\sqrt{\frac{2\bar{\varepsilon}}{L}} \leq 2n\sqrt{\frac{2\bar{\varepsilon}}{L}}\,,\end{aligned} \tag{5.28}$$

where in the second inequality we used part 3 of Lemma A.1 to bound each of the summands of the left-hand side, and in the third inequality we used (5.26). Assume now by contradiction,

$$\operatorname{dist}_1(x^{(s(k'+1))}, A_l) > (g(x^{(s(k'+1))})/\theta)^{1/p}\,.$$

Then by (5.27) applied to $s(k'+1)$ instead of $\bar{k}$, there must necessarily exist $j \neq l$ such that

$$\operatorname{dist}_1(x^{(s(k'+1))}, A_j) \leq (g(x^{(s(k'+1))})/\theta)^{1/p}\,.$$

In particular we have

$$\begin{aligned}\|x^{(s(k'))} - x^{(s(k'+1))}\|_1 &\geq \operatorname{dist}_1(A_l, A_j) - \operatorname{dist}_1(x^{(s(k'+1))}, A_j) - \operatorname{dist}_1(x^{(s(k'))}, A_l) \\ &\geq d - \left(\frac{g(x^{(s(k'))})}{\theta}\right)^{\frac{1}{p}} - \left(\frac{g(x^{(s(k'+1))})}{\theta}\right)^{\frac{1}{p}} \geq d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}}\,,\end{aligned} \tag{5.29}$$

where we used (5.25) in the last inequality. But by the second condition of (5.23), we have

$$d - 2\left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}} > 2n\sqrt{\frac{2\bar{\varepsilon}}{L}}\,. \tag{5.30}$$

Concatenating (5.28), (5.30), and (5.29), we get a contradiction and the claim is proved. An immediate consequence of this claim is $\mathrm{dist}_1(x^{(\tau)}, A_l) < \bar{r}_*$ by (5.27) applied to τ 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 we have that there exists an index $\bar{k} \in [q(\bar{\varepsilon}), q(\bar{\varepsilon}) + n - 1]$ corresponding to a Case 1 or Case 2 step, since there can be at most $n-1$ consecutive Case 3 steps. Second, since by (5.27) we have $\mathrm{dist}_1(x^{(\bar{k})}, A_l) < \bar{r}_*$ and $\bar{k}$ does not correspond to a Case 3 step, by the local identification Theorem 3.3 necessarily $x_i^{(\bar{k})} = 0 \ \forall \ i \in I^c_{A_l}$. Moreover, by the claim every Case 1 and Case 2 step following step $\bar{k}$ happens for points inside $B_1(A_l, \bar{r}_*)$ so it does not change the components corresponding to $I^c_{A_l}$ by the local identification Theorem 3.3. At the same time, Case 3 steps do not increase the support, so that $x_i^{(\bar{k}+l)} = 0$ for every $i \in I^c_{A_l}$, $l \geq 0$. Thus active set identification happens in $\bar{k} \leq q(\bar{\varepsilon}) + n - 1$ steps. ∎

Remark 5.7. When we have an explicit expression for the convergence rate $q(\varepsilon)$, then we can get an active set complexity bound using Theorem 5.6. For instance, we can compare this result with the one for strongly convex objectives, assuming $C = 1, p = 2$, $\theta = u_1/2$, and $f(x^{(k)}) - f(x^{(k+1)}) \leq h_0 q^k$ for some $q \in (0,1)$. These conditions are always satisfied by strongly convex objectives. Applying the theorem we obtain the active set complexity bound

$$q(\bar{\varepsilon}) + n - 1 \leq \left\lceil \max\left(0, \frac{\ln(h_0) - \ln(\min(L, \bar{r}_*^4 u_1^2/16L))}{\ln(1/q)}\right)\right\rceil + n, \tag{5.31}$$

which is always larger than the bound given in (4.5). This is expected, given the weaker assumptions on the convergence of the objective and the weaker (at least in the convex case) error bound.

Remark 5.8. Assume that strict complementarity holds at every stationary point, so that the points in A_i have a common support, for $i \in [1\!:\!C]$; cf. (5.14). Let

$$c_{\min} = \min_{x \in \mathcal{X}^*} \min_{j : x_j \neq 0} x_j \tag{5.32}$$

be the minimal nonzero component of a stationary point. Then the method converges to a set A_l and identifies its support in at most $q(\bar{\varepsilon}) + |I^c_{A_l}|$ iterations, where here the conditions on $\bar{\varepsilon}$ have no explicit dependence on n:

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

with $r(\bar{\varepsilon}) = \left(\frac{2\sqrt{L\bar{\varepsilon}}}{\theta}\right)^{\frac{1}{p}}$ 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 arguments leading to the proof of Theorem 5.6.

5.4. Local active set complexity bound. A key hypothesis to ensure local convergence to a strict local minimum is

$$x^{(k)} \in \mathrm{argmax}\{f(x) \mid x \in \mathrm{conv}(x^{(k)}, x^{(k+1)})\}\,, \tag{5.33}$$

which, in particular, holds when $\alpha_k = \bar{\alpha}_k$ as it is proved in Lemma B.1. The property (5.33) 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 (5.33) holds if and only if $\{f(x^{(k)})\}$ is monotonically nonincreasing.

THEOREM 5.9. *Let x^* be a stationary point which is also a strict local minimizer, isolated from the other stationary points, with value $\tilde{f} = f(x^*)$. Then let β be such that there exists a connected component $V_{x^*,\beta}$ of $f^{-1}((-\infty, \beta])$ satisfying*

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

Then for any $x^{(0)} \in V_{x^*,\beta}$, the sequence $\{x^{(k)}\}$ generated by the AFW with step size $\alpha_k = \bar{\alpha}_k$ converges to x^* and identifies the extended support in at most

$$\left\lceil \max\left(\frac{4(f(x^{(0)}) - \tilde{f})}{\tau}, \frac{8L(f(x^{(0)}) - \tilde{f})}{\tau^2} \right) \right\rceil + n$$

steps with

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

where

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

Proof. As in the proof of Corollary 5.2, the assumptions of Theorem 5.1 are satisfied with $\rho = \frac{1}{2}$. By point 1 of Lemma B.1, the condition $\alpha_k = \bar{\alpha}_k$ on the step sizes implies that $\{x^{(k)}\}$ satisfies (5.33). In particular, $\{x^{(k)}\}$ cannot leave connected components of level sets so that $\{x^{(k)}\} \subset V_{x^*,\beta}$ and

$$\lim_{k\to\infty} f(x^{(k)}) \geq \tilde{f}\,.$$

By (5.7) and (5.9) it follows that

$$f(x^{(0)}) - \tilde{f} \geq [n_1(T) + n_2(T)] \min\left(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2} \right)\,. \tag{5.34}$$

Moreover, applying (5.8) we obtain

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T - n + 1}{2}\,, \tag{5.35}$$

where the second inequality follows from $|S_T| - |S_0| \geq -n + 1$. Concatenating (5.34) and (5.35) we get

$$f(x^{(0)}) - \tilde{f} \geq \frac{T - n + 1}{2} \min\left(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2} \right)\,, \tag{5.36}$$

from which we have the following bound on g_T^*:

$$g_T^* \leq \max\left(\sqrt{\frac{8L(f(x^{(0)}) - \tilde{f})}{T - n + 1}}, \frac{4(f(x^{(0)}) - \tilde{f})}{T - n + 1} \right) \tag{5.37}$$

for $T \geq n$. It is now straightforward to check that if

$$\bar{h} = \left\lceil \max\left(\frac{4(f(x^{(0)}) - \tilde{f})}{\tau}, \frac{8L(f(x^{(0)}) - \tilde{f})}{\tau^2} \right) \right\rceil + n\,,$$

then

$$g^*_{\tilde{h}} < \tau\,.$$

Since (5.34) is derived considering the gap g only in Case 1 and Case 2 indexes, we have that there exists $\tilde{h} \le \bar{h}$ Case 1 or Case 2 index such that $g(x^{(\tilde{h})}) < \tau$. Therefore, by the definition of τ, we get $f(x^{(\tilde{h})}) < m$. We claim that $x^{(h)} \in B_{r_*(x^*)}(x^*)$ for every $h \ge \tilde{h}$. Indeed, since $f(x^{(\tilde{h})}) < m$ and $\{x^{(k)}\}$ cannot leave connected components of level sets we have for every $h \ge \tilde{h}$,

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

where the inclusion follows directly from the definition of m. Since the index $\tilde{h}$ corresponds to a Case 1 or a Case 2 step done in the active set region $B_{r_*(x^*)}(x^*)$ by the local identification Theorem 3.3 the method must have already done all the Case 3 steps needed to identify $I^c(x^*)$. Then we obtain the active set complexity bound

$$\tilde{h} \le \bar{h} = \left\lceil \max\left(\frac{4(f(x^{(0)}) - \tilde{f})}{\tau}, \frac{8L(f(x^{(0)}) - \tilde{f})}{\tau^2} \right) \right\rceil + n\,, \tag{5.38}$$

as desired. ∎

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 extended 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 are finding adaptations for other FW variants and, more generally, for 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.

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

LEMMA A.1. *Given $\{x, y\} \subset \Delta_{n-1}$, $i \in [1:n]$ we have that*

1. $\|e_i - x\| \le \sqrt{2}(e_i - x)_i$ *holds; that*
2. $(y - x)_i \le \|y - x\|_1/2$ *holds; and*
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 $j \ne i$, $(e_i - x)_i = 1 - x_i = \sum_{j\ne i} x_j$. In particular

$$\begin{aligned}\|e_i - x\| &= \left(\sum_{j\ne i} x_j^2 + (e_i - x)_i^2\right)^{\frac{1}{2}} \le \left(\left(\sum_{j\ne i} x_j\right)^2 + (1 - x_i)^2\right)^{\frac{1}{2}} \\ &= \sqrt{2}\left(\sum_{j\ne i} x_j\right) = \sqrt{2}(e_i - x)_i\,.\end{aligned}$$

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

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

and as a consequence,

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

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 $\|d^{(k)}\| \geq \frac{1}{2}\|d^{(k)}\|_1$. We have

$$\|d^{(k)}\| \geq 1 - x_i^{(k)} = \frac{1}{2}(1-x_i^{(k)}) + \sum_{j\neq i} x_j^{(k)} = \frac{1}{2}\|d^{(k)}\|_1\,,$$

where in the first equality we used $\sum_{i=1}^n x_i^{(k)} = 1$ (so that $1 - x_i^{(k)} = \sum_{j\neq i} x_j^{(k)}$) and in the second equality we used $0 \leq x^{(k)} \leq 1$. ∎

Appendix B. Technical results related to step sizes. We now prove several properties related to the step size given in (4.2).

LEMMA B.1. *Consider a sequence $\{x^{(k)}\}$ in Δ_{n-1} such that $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$ with $\alpha_k \in \mathbb{R}_{\geq 0}$, $d^{(k)} \in \mathbb{R}^n$. 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 we have the following:*

1. *If $0 \leq \alpha_k \leq 2p_k/(\|d^{(k)}\|^2 L)$, the sequence $\{x^{(k)}\}$ has the property (5.33).*
2. *If $\alpha_k = \bar{\alpha}_k$, then (5.3) is satisfied with $\rho = \frac{1}{2}$. Additionally, we have*

$$f(x^{(k)}) - f(x^{(k+1)}) \geq L\frac{\|x^{(k+1)} - x^{(k)}\|^2}{2}\,. \tag{B.1}$$

3. *If α_k is given by exact line search, then $\alpha_k \geq \bar{\alpha}_k$ and (5.3) is again satisfied with $\rho = \frac{1}{2}$.*

If $\alpha_k \leq \alpha_k^{\max}$, the condition of point 1 implies $0 \leq \alpha_k \leq 2\bar{\alpha}_k$.

Proof. By the standard descent lemma [7, Proposition 6.1.2], we have

$$f(x^{(k)}) - f(x^{(k)} + \alpha d^{(k)}) \geq \alpha p_k - \alpha^2 \frac{L\|d^{(k)}\|^2}{2}\,. \tag{B.2}$$

It is immediate to check

$$\alpha \nabla f(x^{(k)})^\top d^{(k)} + \alpha^2 \frac{L\|d^{(k)}\|^2}{2} \leq 0 \tag{B.3}$$

for every $0 \leq \alpha \leq \frac{2p_k}{L\|d^{(k)}\|^2}$. Furthermore,

$$\alpha p_k - \alpha^2 \frac{L\|d^{(k)}\|^2}{2} \geq \alpha p_k/2 \geq \alpha^2 \frac{L\|d^{(k)}\|^2}{2} \tag{B.4}$$

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 (B.2) in the first inequality and (B.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 can apply (B.4) since $0 \le \bar\alpha_k \le \frac{p_k}{L\|d^{(k)}\|^2}$. Again by (B.4)

$$f(x^{(k)}) - f(x^{(k+1)}) = f(x^{(k)}) - f(x^{(k)} + \bar\alpha_k d^{(k)}) \ge \bar\alpha_k^2 \frac{L\|d^{(k)}\|^2}{2} = L\frac{\|x^{(k)} - x^{(k+1)}\|^2}{2}\,.$$

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

$$0 = \frac{\partial}{\partial\alpha} f(x^{(k)} + \alpha d^{(k)})|_{\alpha=\alpha_k} = \nabla f(x^{(k)} + \alpha_k d^{(k)})^\top d^{(k)}, \tag{B.5}$$

and therefore,

$$\begin{aligned} -\nabla f(x^{(k)})^\top d^{(k)} &= -\nabla f(x^{(k)})^\top d^{(k)} + \nabla f(x^{(k)} + \alpha_k d^{(k)})^\top d^{(k)} \\ &= -(\nabla f(x^{(k)}) - \nabla f(x^{(k)} + \alpha_k d^{(k)}))^\top d^{(k)} \\ &\le L\|d^{(k)}\|\|x^{(k)} - (x^{(k)} + \alpha_k d^{(k)})\| \\ &= \alpha_k L\|d^{(k)}\|^2\,, \end{aligned} \tag{B.6}$$

where we used (B.5) in the first equality and the Lipschitz condition in the inequality. From (B.6) it follows that

$$\alpha_k \ge \frac{-\nabla f(x^{(k)})^\top d^{(k)}}{L\|d^{(k)}\|^2} \ge \bar\alpha_k,$$

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

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \ge f(x^{(k)}) - f(x^{(k)} + \bar\alpha_k d^{(k)}) \ge \frac{\bar\alpha_k}{2} p_k\,,$$

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

Corollary B.2. *Under the hypotheses of Lemma* B.1*, 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 (B.1) we have

$$f(x^{(k(j))}) - f(x^{(k(j)+1)}) \ge \frac{L}{2}\|x^{(k(j))} - x^{(k(j)+1)}\|^2,$$

and the conclusion follows by monotonicity and boundedness. ∎

We now briefly recall the Armijo line search and the Wolfe conditions with a couple of adaptations to our setting. For the Armijo search we impose the usual condition of sufficient decrease

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \ge c_1 \alpha_k p_k\,, \tag{B.7}$$

and assume that the tentative step sizes are given by $\beta_k^{(0)} = \alpha_k^{\max}$, $\beta_k^{(j+1)} = \gamma\beta_k^{(j)}$ for $c_1, \gamma \in (0,1)$.

LEMMA B.3. *If α_k is determined by the Armijo line search described above, then*

$$\alpha_k \geq \min\left(\alpha_k^{\max}, 2\gamma(1-c_1)\frac{p_k}{L\|d^{(k)}\|^2}\right) \geq \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k \tag{B.8}$$

with $\bar{\alpha}_k = \min(\alpha_k^{\max}, \frac{p_k}{L\|d^{(k)}\|^2})$ *as in* (4.2)*, and* (5.3) *holds with* $\rho = c_1 \min\{1, 2\gamma(1-c_1)\} < 1$.

Proof. From the upper bound on f given in (B.2) it follows that

$$f(x^{(k)}) - f(x^{(k)} + \alpha d^{(k)}) \geq c_1 \alpha p_k \quad \text{for } \alpha \in \left[0, 2(1-c_1)\frac{p_k}{L\|d^{(k)}\|^2}\right] \tag{B.9}$$

and

$$\alpha_k > 2\gamma(1-c_1)\frac{p_k}{L\|d^{(k)}\|^2}\,.$$

Therefore,

$$\alpha_k \geq \min\left(\alpha_k^{\max}, 2\gamma(1-c_1)\frac{p_k}{L\|d^{(k)}\|^2}\right) \geq \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k\,, \tag{B.10}$$

which proves (B.8). We also have

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \geq c_1 \alpha_k p_k \geq c_1 \min\{1, 2\gamma(1-c_1)\}\bar{\alpha}_k p_k\,, \tag{B.11}$$

where we used the Armijo condition (B.7) in the first inequality and (B.8) in the second. Hence, by $c_1, \gamma \in (0,1)$ and $c_1(1-c_1) \leq \frac{1}{4}$, we get that (5.3) holds with $\rho = c_1 \min\{1, 2\gamma(1-c_1)\} < 1$. ∎

The weak Wolfe conditions [33] are (B.7) together with

$$-\nabla f(x^{(k)} + \alpha_k d^{(k)})^\top d^{(k)} \leq c_2 p_k \tag{B.12}$$

for some $c_2 \in (c_1, 1)$.

LEMMA B.4. *Assume* $\alpha_k = \min(\alpha_k^{\max}, \tilde{\alpha}_k)$ *with* $\tilde{\alpha}_k$ *satisfying the weak Wolfe conditions. Then*

$$\alpha_k \geq \min\left(\alpha_k^{\max}, (1-c_2)\frac{p_k}{L\|d^{(k)}\|^2}\right) \geq (1-c_2)\bar{\alpha}_k \tag{B.13}$$

and (5.3) *holds with* $\rho = c_1(1-c_2) < 1$.

Proof. **Case (a)**. $\alpha_k = \alpha_k^{\max}$. Then, trivially, $\alpha_k \geq \bar{\alpha}_k$ and by point 2 of Lemma B.1, (5.3) is satisfied with $\rho = \frac{1}{2}$.

Case (b). The second weak Wolfe condition holds. We have

$$\begin{aligned} c_2 p_k \geq -\nabla f(x^{(k)} + \alpha_k d^{(k)})^\top d^{(k)} &= (-\nabla f(x^{(k)}) + (\nabla f(x^{(k)}) \\ &\quad - \nabla f(x^{(k)} + \alpha_k d^{(k)})))^\top d^{(k)} \\ &\geq p_k - \alpha_k L\|d^{(k)}\|^2\,, \end{aligned} \tag{B.14}$$

where we used (B.12) in the first inequality. Rearranging (B.14) we obtain

$$\alpha_k \geq \frac{(1-c_2)p_k}{L\|d^{(k)}\|^2}\,. \tag{B.15}$$

As for part 1 we can now use the Armijo condition (B.7) to obtain (5.3) with $\rho = c_1(1-c_2)$:

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \geq c_1 \alpha_k p_k \geq c_1(1-c_2)\bar{\alpha}_k p_k \,, \tag{B.16}$$

where we used (B.15) in the second inequality. To conclude, since $\frac{1}{4} \geq c_1(1-c_1) > c_1(1-c_2)$ for $0 < c_1 < c_2 < 1$, the bound (5.3) holds in both cases with $\rho = c_1(1-c_2)$. ∎

Appendix C. 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. Even though rewriting an optimization problem on the simplex can lead to a dramatic increase in complexity, this equivalence is still useful because it allows us to extend the properties we proved on the simplex to generic polytopes. Furthermore, in practice the AFW only needs a linear minimization oracle and the points appearing in the convex combination giving the current iterate [31], while knowledge of the whole transformation between the polytope and the simplex is not needed.

Let P be a polytope and $f : P \to \mathbb{R}^n$ be a function with gradient having Lipschitz constant L. 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 generalize Theorem 4.3, 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 [14] building on [13], 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 λ to polytopes.

To define the AFW algorithm we need a finite set of atoms $\mathcal{A}$ such that $\operatorname{conv}(\mathcal{A}) = P$. 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

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

Finally, let 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 $Ae_i = A^i \in \mathcal{A}$ (but the same results hold with the same proofs if we have an affine transformation $e_i \to Ae_i + b$).

In order to apply Theorem 3.3 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 λ 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 λ: 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

$$\begin{aligned} P^*(y^*) &= \{y \in P \mid \nabla f(y^*)^\top y = \nabla f(y^*)^\top y^*\} \\ &= \operatorname{argmax}\{-\nabla f(y^*)^\top y \mid y \in P\} = \mathcal{F}(-\nabla f(y^*)) \end{aligned}$$

be the face of the polytope P exposed by $-\nabla f(y^*)$. We also define

$$I(y^*) = \{a \in \mathcal{A} \mid \lambda_a(y^*) = 0\}, \quad I^c(y^*) = \mathcal{A} \setminus I(y^*)\,.$$

The complementarity conditions for the generalized multiplier function λ 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 $I(y^*)$,

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

or in other words since $\lambda_{A^i}(y^*) = 0$ if and only if $i \in I(x^*)$,

$$P^*(y^*) = \operatorname{conv}(\{a \in \mathcal{A} \mid a = A^i,\ i \in I(x^*)\})\,. \tag{C.1}$$

A consequence of (C.1) is that given any subset B of P with the SIP, 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 4.3 in slightly different terms.

THEOREM C.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 step sizes 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. The proof follows from Theorem 4.3 and the affine invariance properties discussed above. ∎

In Theorem C.1, 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 B.1 if the AFW uses exact line search, the step size $\bar{\alpha}_k$ (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 4.6, is not maintained by affine transformations, so that instead we have to use a weaker error bound condition. As a possible alternative, in [31] linear convergence of the AFW is proved with

dependence only on affine invariant parameters, so that any version of Theorem 3.3 and Corollary 4.6 depending on those parameters instead of u_1, L would not need this additional analysis.

Let y^* be the unique minimizer of f on P, and let $r_*(y^*) = r_*(x)$ for any x such that $Ax = y^*$, where by the invariance of multipliers $r_*(y^*)$ is well defined. Then let $u > 0$ be the strong convexity constant of f, so that

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

The function $\tilde{f}$ inherits the error bound condition necessary for Corollary 4.6 from the strong convexity of f: for every $x \in \Delta_{|\mathcal{A}|-1}$ by [4, Lemma 2.2], we have

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

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

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

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

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

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

Corollary C.2. *The sequence $\{y^{(k)}\}$ generated by the AFW is in $P^*(y^*)$ for*

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

where $q \in (0,1)$ is the constant related to the linear convergence rate $f(y^{(k)}) - f(y^*) \leq q^k(f(y^{(0)}) - f(y^*))$, $u_A = \frac{u}{2n\theta_A^2}$.

Proof. Let $I = \{i \in [1 : |\mathcal{A}|] \mid A^i \in I(y^*)\}$, $P^* = P^*(y^*)$. Since $P^* = \operatorname{conv}(\mathcal{A} \cap P^*)$ and by (C.1) $\operatorname{conv}(\mathcal{A} \cap P^*) = \operatorname{conv}(\{A^i \mid i \in I\})$, the theorem is equivalent to proving 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 generated 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, $x_i^{(k)} = 0$ for every $i \in I^c$.

Reasoning as in Corollary 4.6 we get that $\operatorname{dist}_1(x^{(k)}, \mathcal{X}^*) < r_*(y^*)$ for every

$$k \geq \frac{\ln(h_0) - \ln(u_A r_*(y^*)^2/2)}{\ln(1/q)} . \tag{C.3}$$

Let $\bar{k}$ be the minimum index such that (C.3) holds. For every $k \geq \bar{k}$ there exists $x^* \in \mathcal{X}^*$ with $\|x^{(k)} - x^*\|_1 < r_*(y^*)$. But $\lambda_i(x) = \lambda_{A^i}(y^*)$ for every $x \in \mathcal{X}^*$ by the invariance of λ, so that we can apply Theorem 3.3 with fixed point x^* and obtain that if $J_k = \{i \in I^c \mid x_i^{(k)} > 0\}$, then $J_{k+1} \leq \max(0, J_k - 1)$. The conclusion follows exactly as in Corollary 4.6. ∎

Acknowledgment. The authors are indebted to three referees for their diligence and constructive suggestions which helped to improve earlier versions of this article.

REFERENCES

[1] F. Bach, *Learning with submodular functions: A convex optimization perspective*, Found. Trends Mach. Learn., 6 (2013), pp. 145–373, https://dl.acm.org/doi/book/10.5555/2602000.

[2] M. V. Balashov, B. T. Polyak, and A. A. Tremba, *Gradient projection and conditional gradient methods for constrained nonconvex minimization*, Numer. Funct. Anal. Optim., 41 (2020), pp. 822–849, https://doi.org/10.1080/01630563.2019.1704780.

[3] L. E. Baum and G. Sell, *Growth transformations for functions on manifolds*, Pacific J. Math., 27 (1968), pp. 211–227, https://doi.org/10.2140/pjm.1968.27.211.

[4] A. Beck and S. Shtern, *Linearly convergent away-step conditional gradient for non-strongly convex functions*, Math. Program., 164 (2017), pp. 1–27, https://doi.org/10.1007/s10107-016-1069-4.

[5] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246, https://doi.org/10.1137/0320018.

[6] D. P. Bertsekas, *Nonlinear Programming*, 3rd ed., Athena Scientific, Belmont, MA, 1999, http://www.athenasc.com/nonlinbook.html (accessed 2020-06-03).

[7] D. P. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, Belmont, MA, 2015, http://www.athenasc.com/convexalgorithms.html (accessed 2020-06-03).

[8] E. G. Birgin and J. M. Martínez, *Large-scale active-set box-constrained optimization method with spectral projected gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125, https://doi.org/10.1023/A:1019928808826.

[9] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Ergeb. Math. Grenzgeb. (3) 36, Springer-Verlag, Berlin, 2013, https://doi.org/10.1007/978-3-662-03718-8.

[10] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Program., 165 (2017), pp. 471–507, https://doi.org/10.1007/s10107-016-1091-6.

[11] I. M. Bomze, F. Rinaldi, and S. Rota Buló, *First-order methods for the impatient: Support identification in finite time with convergent Frank–Wolfe variants*, SIAM J. Optim., 29 (2019), pp. 2211–2226, https://doi.org/10.1137/18M1206953.

[12] J. Burke, *On the identification of active constraints* II*: The nonconvex case*, SIAM J. Numer. Anal., 27 (1990), pp. 1081–1102, https://doi.org/10.1137/0727064.

[13] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211, https://doi.org/10.1137/0725068.

[14] J. V. Burke and J. J. Moré, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595, https://doi.org/10.1137/0804032.

[15] M. D. Canon and C. D. Cullum, *A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm*, SIAM J. Control, 6 (1968), pp. 509–516, https://doi.org/10.1137/0306032.

[16] K. L. Clarkson, *Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm*, ACM Trans. Algorithms, 6 (2010), 63, https://doi.org/10.1145/1824777.1824783.

[17] A. Cristofari, M. Santis, S. Lucidi, and F. Rinaldi, *An active-set algorithmic framework for non-convex optimization problems over the simplex*, Comput. Optim. Appl., 77 (2020), pp. 57–89, https://doi.org/10.1007/s10589-020-00195-x.

[18] M. De Santis, G. Di Pillo, and S. Lucidi, *An active set feasible method for large-scale minimization problems with bound constraints*, Comput. Optim. Appl., 53 (2012), pp. 395–423, https://doi.org/10.1007/s10589-012-9506-7.

[19] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95–110, https://doi.org/10.1002/nav.3800030109.

[20] P. Grigas, A. Lobos, and N. Vermeersch, *Stochastic In-face Frank-Wolfe Methods for Non-convex Optimization and Sparse Neural Network Training*, preprint, https://arxiv.org/abs/1906.03580, 2019.

[21] J. Guelat and P. Marcotte, *Some comments on Wolfe's 'away step,'* Math. Program., 35 (1986), pp. 110–119, https://doi.org/10.1007/BF01589445.

[22] W. W. Hager, D. T. Phan, and H. Zhang, *Gradient-based methods for sparse recovery*, SIAM J. Imaging Sci., 4 (2011), pp. 146–165, https://doi.org/10.1137/090775063.

[23] W. W. Hager and H. Zhang, *A new active set algorithm for box constrained optimization*, SIAM J. Optim., 17 (2006), pp. 526–557, https://doi.org/10.1137/050635225.

[24] W. W. Hager and H. Zhang, *An active set algorithm for nonlinear optimization with polyhedral constraints*, Sci. China Math., 59 (2016), pp. 1525–1542, https://doi.org/10.1007/s11425-016-0300-6.

[25] W. L. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, J. Convex Anal., 11 (2004), pp. 251–266, http://www.heldermann.de/JCA/JCA11/jca11017.htm (accessed 2020-06-03).

[26] A. N. Iusem, *On the convergence properties of the projected gradient method for convex optimization*, Comput. Appl. Math., 22 (2003), pp. 37–52, http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-03022003000100003&nrm=iso (accessed 2020-06-03).

[27] M. Jaggi, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, in ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, JMLR.org, 2013, p. I-427–I-435, https://dl.acm.org/doi/10.5555/3042817.3042867.

[28] T. Kerdreux, A. d'Aspremont, and S. Pokutta, *Restarting Frank-Wolfe*, in Proceedings of Machine Learning Research (PMLR), K. Chaudhuri and M. Sugiyama, eds., Proceedings of Machine Learning Research 89, PMLR, 2019, pp. 1275–1283, http://proceedings.mlr.press/v89/kerdreux19a.html (accessed 2020-06-03).

[29] R. G. Krishnan, S. Lacoste-Julien, and D. Sontag, *Barrier Frank-Wolfe for marginal inference*, in Advances in Neural Information Processing Systems, 2015, pp. 532–540, https://dl.acm.org/doi/10.5555/2969239.2969299.

[30] S. Lacoste-Julien, *Convergence Rate of Frank-Wolfe for Non-convex Objectives*, preprint, https://arxiv.org/abs/1607.00345, 2016.

[31] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of Frank-Wolfe optimization variants*, in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, MIT Press, Cambridge, MA, 2015, pp. 496–504, https://dl.acm.org/doi/10.5555/2969239.2969295.

[32] Y. Nesterov, *Lectures on Convex Optimization*, Springer Optim. Appl. 137, Springer, Cham, 2018, https://doi.org/10.1007%2F978-3-319-91578-4.

[33] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.

[34] J. Nutini, I. Laradji, and M. Schmidt, *Let's Make Block Coordinate Descent Go Fast: Faster Greedy Rules, Message-Passing, Active-Set Complexity, and Superlinear Convergence*, preprint, https://arxiv.org/abs/1712.08859, 2017.

[35] J. Nutini, M. Schmidt, and W. Hare, *"Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?*, Optim. Lett., 13 (2019), pp. 645–655, https://doi.org/10.1007/s11590-018-1325-z.

[36] J. Peña and D. Rodriguez, *Polytope conditioning and linear convergence of the Frank–Wolfe algorithm*, Math. Oper. Res., 44 (2019), pp. 1–18, https://doi.org/10.1287/moor.2017.0910.

[37] Y. Sun, H. Jeong, J. Nutini, and M. Schmidt, *Are we there yet? Manifold identification of gradient-related proximal methods*, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1110–1119, http://proceedings.mlr.press/v89/sun19a.html (accessed 2020-06-03).

[38] P. Wolfe, *Convergence theory in nonlinear programming*, in Integer and Nonlinear Programming, North-Holland, Amsterdam, 1970, pp. 1–36.

[39] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079, https://doi.org/10.1137/0331048.

[40] Y. Xu and T. Yang, *Frank-Wolfe Method is Automatically Adaptive to Error Bound Condition*, preprint, https://arxiv.org/abs/1810.04765, 2018.