FIRST-ORDER METHODS FOR THE IMPATIENT: SUPPORT IDENTIFICATION IN FINITE TIME WITH CONVERGENT FRANK-WOLFE VARIANTS

IMMANUEL M. BOMZE*, FRANCESCO RINALDI†, AND SAMUEL ROTA BULÒ‡

Abstract. In this paper, we focus on the problem of minimizing a non-convex function over the unit simplex. We analyze two well-known and widely used variants of the Frank-Wolfe algorithm and first prove global convergence of the iterates to stationary points both when using exact and Armijo line search. Then we show that the algorithms identify the support in a finite number of iterations (the identification result does not hold for the classic Frank-Wolfe algorithm). This, to the best of our knowledge, is the first time a manifold identification property has been shown for such a class of methods.

Key words. Surface Identification, Manifold Identification, Active Set, Finite Convergence

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

1. Introduction. The minimization of a (possibly non-convex) function over the probability simplex is a problem arising in many different contexts like, e.g., machine learning, statistics and economics (see, e.g., [7, 11] for an overview of real-world applications). When dealing with this kind of problems, Frank-Wolfe variants (see, e.g., [16] and references therein) guarantee good scalability, thanks to the way they handle the feasible set, and also give a sparse representation of the iterates, thus offering a good alternative to projected gradient algorithms. Anyway, some may argue that projected gradient methods still represent the best choice in the considered framework, since they can identify the sparsity pattern, i.e., the final set of non-zero variables, in a finite number of iterations (under some specific assumptions). This feature is particularly useful if the solution of the problem is sparse and we just want to find its support, since it means we do not need to run the algorithm until convergence. It is also important when trying to speed-up a given algorithm. Indeed, after we identify the set of non-zero variables, we could simply apply some more sophisticated Newton-like method over the lower-dimensional space those variables describe. Such a feature may also help to develop suitable support identification/active-set strategies, like the ones described in, e.g., [2, 4, 9, 10, 12, 14].

There exists a considerable number of papers analyzing support/active-set identification properties of optimization methods. Bertsekas first showed in [1] that the projected gradient method identifies the sparsity pattern in a finite number of iterations when using non-negativity constraints. In [6] the authors showed that some simple algorithms (including projected gradient) would, in a finite number of iterations, identify the face of a polyhedral feasible region on which the solutions to an optimization problem occur. These results where generalized in [24] to the case of non-polyhedral convex sets. Analysis for nonconvex constraints is reported in [5, 15].

*ISOR, VCOR & ds:UniVie, Universität Wien, Austria (immanuel.bomze@univie.ac.at)
†Dipartimento di Matematica, Università di Padova, Italy (rinaldi@math.unipd.it)
‡Mapillary Research, Graz, Austria (samuel@mapillary.com)

The support identification property has also been established for other algorithms like certain coordinate descent and stochastic gradient methods [18, 25], proximal gradient methods (see, e.g., [19, 21]) and sequential minimal optimization methods for SVM training [22]. In [7], the problem of minimizing a convex function over the probability simplex is considered, and coreset-based results are reported for fully corrective versions of some Frank-Wolfe variants. Recall that a coreset is a face of the simplex with the property that the minimum of the function on the face is a good approximate solution of the full problem. It is further important to remark that fully corrective algorithms heavily rely on the fact that a minimum of the function over a given face can be calculated at each iteration. Hence, those algorithms cannot be considered when dealing with non-convex problems.

In the present paper, we consider two well-known variants of the Frank-Wolfe algorithm, namely away-step Frank-Wolfe [23] and pairwise Frank-Wolfe [16, 20], and prove global convergence of their iterates to stationary points when using exact or Armijo line search (in the sense of characterizing all accumulation points of iterates by stationarity), and moreover global convergence for the full iteration sequence for the away-step variant. These results then enable us to prove support identification in a finite number of iterations for those algorithms. More specifically, when considering a convergent sequence (x^k) generated by one of those Frank-Wolfe variants, we have that it converges to a stationary point $\bar{x}$. Furthermore, we can be sure the iterates x^k will match the sparsity pattern of $\bar{x}$ when k is sufficiently large (if strict complementarity holds at $\bar{x}$). This, to the best of our knowledge, is the first time that a support identification result is proved for Frank-Wolfe like algorithms.

This result is quite surprising if we take into account the fact that the classic Frank-Wolfe algorithm does not guarantee support identification in finite time. We will give some examples later on (see Section 4) where all iterates generated by the algorithm have full support (i.e., number of nonzero coordinates equal to the number of variables in the problem), and the limit point of the iterate sequence does not.

The paper is organized as follows. After a preliminary analysis of the problem in Section 2, we describe in depth the algorithmic framework in Section 3. In Section 4 we establish global convergence and support identification property of the methods. Finally, in Section 5, we draw some conclusions.

2. Preliminary Analysis of the Problem. Denoting by $e = (1, \dots, 1)^\top$ the n-dimensional vector with all entries equal to one, the problem we consider here is the following:

$$\min_{x \in \Delta} \quad f(x) \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $\Delta = \{x \in \mathbb{R}^n : e^\top x = 1, x \geq 0\}$ is the probability simplex. A class of C^2-objective functions f including all quadratic functions will be considered in this paper. For any fixed $x \in \Delta$ and any feasible direction d (we will construct d such that always $[0, 1] \subseteq I_{\text{feas}}(x, d) := \{\alpha \in \mathbb{R} : x + \alpha d \in \Delta\}$ holds), denote by

$$\varphi_d^x(\alpha) = f(x + \alpha d)\,, \quad \alpha \in I_{\text{feas}}(x, d)$$

with derivatives $\dot{\varphi}_d^x(\alpha) = d^\top \nabla f(x + \alpha d)$ and $\ddot{\varphi}_d^x(\alpha) = d^\top \nabla^2 f(x + \alpha d) d$.

We now give a key assumption on curvature of φ_d^x that will be needed to prove convergence of iterates (see Subsection 4.2). As we will see later on, this will guarantee that iteration is *homotopical* for the algorithms we analyze in the paper.

ASSUMPTION 2.1. *Any φ_d^x is either concave or strictly convex over $I_{\text{feas}}(x,d)$. Furthermore, curvature should be bounded away from zero in the strictly convex case along descent directions: to be more precise, for all $x \in \Delta$ and all d with non-concave φ_d^x we ask existence of $\eta_d > 0$ such that*

$$\eta_d \le \ddot{\varphi}_d^x(\alpha) \quad \text{for all } \alpha \in I_{\text{feas}}(x,d)\,, \quad \text{if} \quad \dot{\varphi}_d^x(0) < 0\,. \tag{2.2}$$

All quadratic functions $f(x) = x^\top Q x + c^\top x$, where Q is a possibly indefinite symmetric matrix, satisfy (2.2) with $\eta_d = d^\top Q d$ for all $x \in \Delta$. But many more functions f may meet Assumption 2.1, for example[1] the function $f(x) = c^\top x + \sqrt{x^\top Q x}$ for indefinite but strictly (co)positive Q (similar functions are used in volatility modeling). Then

$$[(x+\alpha d)^\top Q (x+\alpha d)]^{3/2} \ddot{\varphi}_d^x(\alpha) = (x^\top Q x)(d^\top Q d) - (d^\top Q x)^2$$

does not depend on α but can change sign with varying d.

For proving global convergence of the methods and support identification results (see Section 4), we need an essential global estimate following by continuity of $\nabla^2 f$ over Δ (a set of diameter $\sqrt{2}$) made explicit in the following observation:

OBSERVATION 2.1. *For all directions d with $\|d\| \le \sqrt{2}$ and all $\alpha \in I_{\text{feas}}(x,d)$ we have bounded curvatures $\ddot{\varphi}_d^x(\alpha)$, or slightly more general $\|\nabla^2 f(x)\|_{\text{spec}} \le K$ for all $x \in \Delta$ with the spectral norm $\|.\|_{\text{spec}}$, implying*

$$|\,\ddot{\varphi}_d^x(\alpha)\,| = |\,d^\top \nabla^2 f(x+\alpha d) d\,| \le 2K \quad \text{for all } x \in \Delta\,, \ \text{if } \|d\| \le \sqrt{2}\,. \tag{2.3}$$

We further notice that minimizing a function $h(x)$ over a polytope P can be written as Problem (2.1). Let $V = \begin{bmatrix} v_1, & \dots, & v_m \end{bmatrix} \in \mathbb{R}^{n \times m}$ be the matrix whose columns are the vertices of P. Since any point $y \in P$ can be expressed as a convex combination of the columns of V, the problem $\min\{h(y) : y \in P\}$ can be rewritten as the problem $\min\{f(x) = h(Vx) : x \in \Delta\}$. We note that

1. $\bar{x}$ is a stationary point for f over Δ, cf. (3.1) below, if and only if $\bar{y} = V\bar{x}$ is a stationary point for h over P, i.e., satisfies the KKT conditions;
2. d is a descent direction for f at $x \in \Delta$ if Vd is one for h at $y = Vx \in P$; and
3. condition (2.2) carries over from h to f too, as $\nabla^2 f(x) = V^\top \nabla^2 h(Vx) V$.

3. Frank-Wolfe Variants for Minimization over the Simplex. In this section, we describe two well-known Frank-Wolfe variants that can be used to minimize a function over the probability simplex. In order to do that, we report below the generic scheme related to those iterative algorithms (see Algorithm 3.1). Beforehand we recall that $x^* \in \Delta$ is a stationary point for the problem (2.1) if and only if

$$\nabla_r f(x^*) \ge \nabla f(x^*)^\top x^* \quad \text{for all } r \text{ with equality if } x_r^* > 0\,. \tag{3.1}$$

By construction, either the algorithm stops after finitely many iterations at a stationary point, or else the generated sequence takes infinitely many values in Δ as $f(x^{k+1}) < f(x^k)$.

[1] We are grateful to Werner Schachinger who pointed to this in personal communication.

Algorithm 3.1 `Line Search Algorithmic Scheme`

1 Choose a point $x^0 \in \Delta$
2 For $k = 0, 1, \ldots$
3 If x^k is a stationary point (3.1), then STOP
4 Compute a feasible descent direction d^k at x^k
5 Compute a stepsize $\alpha_k \in (0, 1]$ via line search for improving the objective
6 Set $x^{k+1} = x^k + \alpha_k d^k$
7 End for

3.1. Frank-Wolfe type directions. At every iteration k of Algorithm 3.1, we compute, at Step 4, a feasible descent search direction d^k that is used to generate the new iterate x^{k+1}. We describe here the different type of directions that can be used in Algorithm 3.1. We indicate the set of all indices related to the coordinates of vector x by $I = \{1, \ldots, n\}$, and by $S_k = \{i \in I : \ x_i^k > 0\}$ we denote the support of x^k.

The Frank-Wolfe and the away-step directions (see, e.g., [13, 16]), computed in x^k are respectively:

$$d_{FW}^{x^k} = e_{\hat{\imath}} - x^k, \quad \hat{\imath} \in \operatorname*{Argmin}_{i \in I} \{\nabla_i f(x^k)\}\,; \tag{3.2}$$

$$d_A^{x^k} = x^k - e_{\hat{\jmath}}, \quad \hat{\jmath} \in \operatorname*{Argmax}_{j \in S_k} \{\nabla_j f(x^k)\}\,. \tag{3.3}$$

We further indicate with $x_{\hat{\jmath}}^k$ the $\hat{\jmath}$-th coordinate of x^k, where $\hat{\jmath}$ is defined as in (3.3). Taking into account (3.2) and (3.3), we consider the following two search directions:

(AFW) Away-step Frank-Wolfe direction:

$$d_{AFW}^{x^k} = \begin{cases} d_{FW}^{x^k}\,, & \text{if } \nabla f(x^k)^\top d_{FW}^{x^k} \le \nabla f(x^k)^\top d_A^{x^k}\,, \\ \dfrac{x_{\hat{\jmath}}^k}{1 - x_{\hat{\jmath}}^k}\, d_A^{x^k}\,, & \text{otherwise.} \end{cases}$$

(PFW) Pairwise Frank-Wolfe direction:

$$d_{PFW}^{x^k} = x_{\hat{\jmath}}^k (d_{FW}^{x^k} + d_A^{x^k}) = x_{\hat{\jmath}}^k (e_{\hat{\imath}} - e_{\hat{\jmath}})\,,$$

where $\hat{\imath}$ and $\hat{\jmath}$ are defined as in (3.2) and (3.3), respectively.

It is easy to verify that all above directions are strict descent directions, i.e., satisfy $\dot{\varphi}_d^x(0) = \nabla f(x)^\top d < 0$.

3.2. Computation of the stepsize. In the framework of Algorithm 3.1, given $x \in \Delta$ and a descent direction d at x, we aim at the largest (global) minimizer $\alpha_d^x > 0$ of $\varphi_d^x(\alpha)$ over $(0, 1]$, i.e.

$$\alpha_d^x := \max \operatorname*{Argmin}_{\alpha \in (0,1]} \varphi_d^x(\alpha)\,. \tag{3.4}$$

Obviously, any global interior minimizer of φ in $(0,1)$ satisfies the first-order condition

$$0 = \dot{\varphi}_d^x(\alpha_d^x) = \dot{\varphi}_d^x(0) + \alpha_d^x \ddot{\varphi}_d^x(\tilde{\alpha})$$

for some $\tilde{\alpha} \in [0,1]$ depending on d and x. Hence, if $\ddot{\varphi}_d^x(\tilde{\alpha}) > 0$ we have

$$\alpha_d^x = \frac{-\dot{\varphi}_d^x(0)}{\ddot{\varphi}_d^x(\tilde{\alpha})}\,. \tag{3.5}$$

3.2.1. Exact and Armijo's line search. Exact line search chooses, at a given iteration k, the largest minimizer of $\varphi_{d^k}^{x^k}(\alpha)$ over $(0,1]$, that is

$$\alpha_k := \alpha_{d^k}^{x^k} \quad \text{defined as in (3.4) for } x = x^k \text{ and } d = d^k\,. \tag{3.6}$$

Unless the function $\varphi_{d^k}^{x^k}$ has some special structure (e.g., convexity/concavity), determining the step size in (3.6) might be in general an expensive task. More practical strategies perform an inexact line search to identify the stepsize giving sufficient reductions in the objective function at a minimal cost. A classic example is represented by the Armijo method (see, e.g., [3] and references therein). This method iteratively shrinks the step size in order to guarantee a sufficient reduction of the objective function. It represents a good way to replace exact line search in cases when it gets too costly. In practice, we fix parameters $\delta \in (0,1)$ and $\gamma \in (0, \frac{1}{2})$, and start with maximal feasible stepsize equal to one. We then try steps $\alpha = \delta^m$ with $m \in \mathbb{N}_0 = \{0,1,2,\ldots\}$ until the sufficient decrease inequality

$$f(x^k + \alpha d^k) \le f(x^k) + \gamma\alpha\nabla f(x^k)^\top d^k \tag{3.7}$$

is satisfied, i.e., choose

$$m(x^k, d^k) := \min\{m \in \mathbb{N}_0 : (3.7) \text{ is satisfied for } \alpha = \delta^m\} < \infty$$

and set

$$\alpha_k = \delta^{m(x^k,d^k)} \quad \text{as well as} \quad x^{k+1} = x^k + \alpha_k d^k\,. \tag{3.8}$$

Observe that under Assumption 2.1 on the curvature of φ_d^x, all stepsize variants we discuss here have in common that always a full feasible step is taken unless in the case of strictly convex φ_d^x where $\ddot{\varphi}_d^x(\alpha) > 0$ for all $\alpha \in [0,1]$. So $\alpha_k < 1$ is possible only if $\ddot{\varphi}_{d^k}^{x^k}(0) > 0$, for any strict descent direction d^k at x^k.

3.2.2. Theoretical properties of line searches. Now we prove that function f reduces when moving from x^k to x^{k+1}, and that the sequence of the directional derivatives along the search direction converges to zero when using the Armijo rule. We will further see that a similar result also holds for the exact line search.

PROPOSITION 3.1. *Let (x^k) be the sequence generated by Algorithm 3.1 using Armijo line search defined in (3.8), with any strict descent direction d^k satisfying $\|d^k\| \le \sqrt{2}$. Then we have*

(a) if $x^{k+1} \ne x^k$, then $f(x^{k+1}) < f(x^k)$;
(b) if $x^{k+1} \ne x^k$ for all $k \in \mathbb{N}$, then $\lim_{k\to\infty} \nabla f(x^k)^\top d^k = 0$.

Proof. We first notice that in a finite number of steps the Armijo line search finds a step satisfying condition (3.7). Then, due to the fact that d^k is such that $\nabla f(x^k)^\top d^k < 0$, we get that

$$f(x^{k+1}) < f(x^k).$$

Using again (3.7), we have

$$f(x^k) - f(x^{k+1}) \geq \gamma \alpha_k |\nabla f(x^k)^\top d^k|. \tag{3.9}$$

Since $f(x^k)$ is monotonically decreasing and bounded in k, we can write

$$\lim_{k\to\infty} \alpha_k |\nabla f(x^k)^\top d^k| = 0. \tag{3.10}$$

Let us consider, by contradiction, that (b) does not hold. In this case, due to the fact that $\{\nabla f(x^k)^\top d^k\}$ is bounded, there exists an infinite subsequence k_j such that

$$\lim_{j\to\infty} \nabla f(x^{k_j})^\top d^{k_j} = -\xi < 0, \tag{3.11}$$

with $\xi > 0$. Considering the limit in (3.10), we need to have

$$\lim_{j\to\infty} \alpha_{k_j} = 0. \tag{3.12}$$

Using compactness of the feasible set Δ, we know that it is possible to get subsequence (for ease of notation we again call it k_j) such that

$$\lim_{k_j\to\infty} x^{k_j} = \widehat{x} \quad \text{and} \quad \lim_{k_j\to\infty} d^{k_j} = \widehat{d}\,. \tag{3.13}$$

Using continuity of the gradient, we thus can write

$$\lim_{j\to\infty} \nabla f(x^{k_j})^\top d^{k_j} = \nabla f(\widehat{x})^\top \widehat{d} = -\xi < 0. \tag{3.14}$$

Taking into account (3.12), we in particular have for k_j sufficiently large

$$\alpha_{k_j} < 1\,.$$

Therefore

$$f(x^{k_j} + \frac{\alpha_{k_j}}{\delta} d^{k_j}) - f(x^{k_j}) > \frac{\gamma \alpha_{k_j}}{\delta} \nabla f(x^{k_j})^\top d^{k_j}\,. \tag{3.15}$$

Using the mean value theorem we can replace the left hand side and write

$$\frac{\alpha_{k_j}}{\delta} \nabla f(y^{k_j})^\top d^{k_j} > \gamma \frac{\alpha_{k_j}}{\delta} \nabla f(x^{k_j})^\top d^{k_j}, \tag{3.16}$$

with $y^{k_j} = x^{k_j} + \theta_{k_j} \frac{\alpha_{k_j}}{\delta} d^{k_j}$ and $\theta_{k_j} \in (0,1)$. Now, dividing by $\frac{\alpha_{k_j}}{\delta} > 0$ and taking into account that $y^{k_j} \to \widehat{x}$ due to (3.12), we have

$$\nabla f(\widehat{x})^\top \widehat{d} \geq \gamma \nabla f(\widehat{x})^\top \widehat{d},$$

which finally gives us

$$\xi \leq \gamma \xi\,,$$

thus contradicting $\gamma < 1$ and proving that (b) holds. ∎

Proposition 3.1 still holds when considering a stepsize $\bar{\alpha}_k \in (0, 1]$ satisfying the following inequality:

$$f(x^k + \bar{\alpha}_k d^k) \leq f(x^k + \alpha_k d^k),$$

where α_k is the stepsize obtained using the Armijo's rule. Indeed, if the above inequality is satisfied, then (3.9) holds as well as the rest of the proof (see also Remark 5 in [10]). Hence, we can easily get the following result:

COROLLARY 3.2. *Let (x^k) be the sequence of points in the feasible set Δ generated by Algorithm 3.1 using exact line search defined in (3.6), with any feasible descent direction d^k. Then we have*

(a) if $x^{k+1} \neq x^k$ then $f(x^{k+1}) < f(x^k)$;
(b) if $x^{k+1} \neq x^k$ for all $k \in \mathbb{N}$, then $\lim_{k\to\infty} \nabla f(x^k)^\top d^k = 0$.

Summarizing Proposition 3.1 and Corollary 3.2, we get under the stepsize choice of (3.6) or (3.8) that

$$\dot{\varphi}_k(0) = \nabla f(x^k)^\top d^k \to 0 \quad \text{as} \quad k \to \infty\,. \tag{3.17}$$

4. Convergence results. To clarify terminology, let us stress we use a common one: "global convergence" means that we establish the stationarity property for *all accumulation points* of the sequence of iterates (x^k), regardless whether or not there may be more than one accumulation point. By contrast, "iterates convergence" means convergence of the full sequence (x^k). Under mild assumptions which generically are true[2] we can show that there is only one accumulation point (which then enjoys stationarity by the global convergence results), if the sequence is generated by the (AFW) rule.

4.1. Global convergence analysis. In this section, for every considered choice of the direction d^k, we establish global convergence to stationary points of the algorithmic scheme described above. Since the arguments for the different stepsize choices vary slightly, we chose to split the treatment. However, the two search direction choices are treated simultaneously, in an attempt of being concise.

THEOREM 4.1. *Let (x^k) be a sequence generated by Algorithm 3.1 where*

- *the search direction d^k is computed according to (AFW) or (PFW) rule;*
- *the stepsize α_k is computed using the Armijo line search described in (3.8).*

Then, either an integer $\bar{k} \geq 0$ exists such that $x^{\bar{k}}$ is a stationary point for problem (2.1), or else the sequence (x^k) is infinite and every limit point x^* of the sequence is a stationary point (3.1) for problem (2.1).

Proof. We first consider the case when the algorithm stops after a finite number of iterations $\bar{k}$. This can only happen if condition at Step 3 of Algorithm 3.1 is satisfied, i.e., if no direction d_{AFW} can be chosen, which is the case if and only if $x^{\bar{k}}$ is a stationary point.
Now we consider the case when the sequence (x^k) is infinite. Arguing by contradiction, assume that there is an i such that $\nabla_i f(x^*) < \nabla f(x^*)^\top x^*$. We again distinguish cases:
Case 1 [not needed for (PFW)]. There is a subsequence k_j along which $x^{k_j} \to x^*$ and

[2] namely that there are only finitely many stationary points of the problem (2.1).

$d^{k_j} = d_{FW}^{x^{k_j}} = e^{i_j} - x^{k_j}$ for all j, where e^i denotes the ith column of the $n \times n$ identity matrix (and $i_j \in \{1, \dots, n\}$ suitably chosen). Then

$$\begin{aligned} \dot\varphi_{k_j}(0) &= \nabla f(x^{k_j})^\top d_{FW}^{x^{k_j}} = \nabla_{i_j} f(x^{k_j}) - \nabla f(x^{k_j})^\top x^{k_j} \\ &\le \nabla_i f(x^{k_j}) - \nabla f(x^{k_j})^\top x^{k_j} \to \nabla_i f(x^*) - \nabla f(x^*)^\top x^* < 0 \end{aligned}$$

as $j \to \infty$, contradicting (3.17).

Case 2a. There is a subsequence k_j along which $x^{k_j} \to x^*$ and such that there is an $\eta > 0$ with

$$x_{r_j}^{k_j} \ge \eta \text{ for all } j\,, \tag{4.1}$$

where in the (AFW) case

$$d^{k_j} = \frac{x_{r_j}^{k_j}}{1 - x_{r_j}^{k_j}}\, d_A^{x^{k_j}} = \frac{x_{r_j}^{k_j}}{1 - x_{r_j}^{k_j}}\, (x^{k_j} - e^{r_j}) \tag{4.2}$$

whereas in the (PFW) case

$$d^{k_j} = x_{r_j}^{k_j}\, (d_{FW}^{x^{k_j}} + d_A^{x^{k_j}}) = x_{r_j}^{k_j}\, (e^{\tilde r_j} - e^{r_j}) \tag{4.3}$$

with $e^{\tilde r_j}$ the Frank-Wolfe vertex and e^{r_j} the away-step vertex. Then in the (AFW) case, as $\frac{x_{r_j}^{k_j}}{1 - x_{r_j}^{k_j}} \ge \frac{\eta}{1-\eta} > 0$ holds for all j, we arrive at

$$\begin{aligned} \tfrac{1-\eta}{\eta}\, \dot\varphi_{k_j}(0) &= \tfrac{1-\eta}{\eta} \nabla f(x^{k_j})^\top d_A^{x^{k_j}} \\ &\le \nabla f(x^{k_j})^\top d_A^{x^{k_j}} \le \nabla f(x^{k_j})^\top d_{FW}^{x^{k_j}} \le \nabla_i f(x^{k_j}) - \nabla f(x^{k_j})^\top x^{k_j} \\ &\to \nabla_i f(x^*) - \nabla f(x^*)^\top x^* < 0 \ \text{ as } j \to \infty\,, \end{aligned}$$

again contradicting (3.17). Similarly in the (PFW) case, the contradiction is obtained via

$$\begin{aligned} \tfrac{1}{\eta}\, \dot\varphi_{k_j}(0) &= \tfrac{1}{\eta}\, \nabla f(x^{k_j})^\top \left(d_{FW}^{x^{k_j}} + d_A^{x^{k_j}} \right) \\ &\le \nabla f(x^{k_j})^\top \left(d_{FW}^{x^{k_j}} + d_A^{x^{k_j}} \right) \le \nabla f(x^{k_j})^\top d_{FW}^{x^{k_j}} \\ &\le \nabla_i f(x^{k_j}) - \nabla f(x^{k_j})^\top x^{k_j} \to \nabla_i f(x^*) - \nabla f(x^*)^\top x^* < 0 \end{aligned}$$

as $j \to \infty$. Hence the only remaining possibility is now

Case 2b. **any** convergent subsequence $x^{k_j} \to x^*$ with limit x^* satisfies

$$x_{r_j}^{k_j} \to 0 \text{ as } j \to \infty\,, \tag{4.4}$$

where eventually (4.2) or (4.3) holds. Irrespective of the chosen direction, at least one such sequence (s_j) exists by assumption that x^* is a limit point of (x^k). Consider

this subsequence and their immediate successors $k_j = s_j + 1$. By (4.2) or (4.3), we know

$$\|x^{s_j+1} - x^{s_j}\| \leq \alpha_{s_j} \max\left\{ \left\| \frac{x_{r_j}^{k_j}}{1 - x_{r_j}^{k_j}} d_A^{x^{k_j}} \right\|, \|x_{r_j}^{k_j} (d_{FW}^{x^{k_j}} + d_A^{x^{k_j}})\| \right\} \leq \sqrt{2} \frac{x_{r_j}^{k_j}}{1 - x_{r_j}^{k_j}} \to 0$$

as $j \to \infty$, since $\|d_A^{x^{k_j}}\| = \|x^{k_j} - e^{r_j}\| \leq \operatorname{diam}\Delta = \sqrt{2}$ and likewise $\|d_{FW}^{x^{k_j}} + d_A^{x^{k_j}}\| = \|e^{\tilde{r}_j} - e^{r_j}\| \leq \operatorname{diam}\Delta = \sqrt{2}$. Therefore also $x^{s_j+1} \to x^*$ as $j \to \infty$, and we may also consider (4.4) with (4.2) or (4.3) for the successor sequence $k_j = s_j + 1$. By suitable thinning (s_j) if necessary, we may and do assume that eventually $r_j = r$ for all j. Then $x_r^{s_j+1} = 0$ eventually holds because otherwise $\alpha_{s_j} < 1$ and Prop. A.2 applies, contradicting (4.4). Applying our conclusion (4.4) with (4.2) or (4.3) now to $k_j = s_j + 1$, we see that also an away-step $x^{s_j+1} - e^h$ (or a PFW step involving e^h as away-step vertex) with $h \neq r$ is selected for $k = s_j + 1$ (if j is large enough) with the property (again, after suitable thinning) that $x_h^{s_j+1} \to 0$ as $j \to \infty$ but still we have, by construction of the away (or PFW) step, $x_r^{s_j+2} = 0$ for all large enough j. So again we have $x^{s_j+2} \to x^*$ as $j \to \infty$, hence an index $t \notin \{r, h\}$ would be chosen for the away step at $k = s_j + 2$, and repeating the argument less than n times, no choice for d_A would be left, which is absurd in view of the fact that the sequence is infinite, whence neither Case 1 nor Case 2a applies. So the theorem is proved. ∎

We close this section by proving global convergence of the Algorithm 3.1 when using the exact line search for calculating the stepsize.

THEOREM 4.2. *Let (x^k) be a sequence generated by Algorithm 3.1 where*

- *the search direction d^k is computed according to (AFW) or (PFW) rule;*
- *the stepsize α_k is computed using the line search described in (3.6).*

Then, either an integer $\bar{k} \geq 0$ exists such that $x^{\bar{k}}$ is a stationary point for problem (2.1), or else the sequence (x^k) is infinite and every limit point x^* of the sequence is a stationary point (3.1) for problem (2.1).

Proof. The proof is very similar to the one given for the Armijo line search. The only difference is in Case 2b where we yield a contradiction by applying Prop. A.1. ∎

4.2. Iterates convergence and support identification in finite time. We start with a general observation, in particular applicable to (AFW) and (PFW) directions. All we need is that the conclusions of Theorems 4.1 and 4.2 hold, namely that all accumulation points are stationary; with this property, any strict local minimizer which is isolated among all stationary points can be shown to attract all sequences generated by Algorithm 3.1 which start close enough to it. Conversely, if the limit point attracts all iterates starting close enough to it, then necessarily this must be an isolated stationary point and a strict local minimizer of f over Δ.

Note that the equivalence below holds irrespective whether or not there are non-strict local solutions to (2.1).

THEOREM 4.3. *Let Assumption 2.1 hold. Consider Algorithm 3.1 with any descent direction and any stepsize, such that all accumulation points of generated sequences (x^k) are stationary. Then the following two statements on a stationary point $p \in \Delta$ are equivalent:*

(a) *there is a p-neighbourhood $U \subseteq \Delta$ with no stationary point in $U \setminus \{p\}$, and $f(x) > f(p)$ for all $x \in U \setminus \{p\}$.*
(b) *there is a p-neighbourhood $V \subseteq \Delta$ such that every sequence (x^k) starting at $x^0 \in V$ converges to p.*

Proof. (a) ⇒ (b): Let $\varepsilon > 0$ be so small that $B := \{x \in \Delta : \|x - p\| \le \varepsilon\} \subseteq U$ and define

$$\sigma := \min\{f(x) : x \in \Delta\,,\ \|x - p\| = \varepsilon\} - f(p) > 0\,.$$

Then $V = \{x \in \Delta : f(x) < f(p) + \sigma\,,\ \|x - p\| < \varepsilon\} \subset U$ is relatively open in Δ and contains p, so a neighbourhood of p in Δ. We claim that any sequence starting in V will remain there forever. Indeed, suppose $x^{k+1} \notin V$ but $x^k \in V$ for some k; then by convexity or concavity of f along $conv(x^k, x^{k+1})$ we have

$$f(\lambda x^{k+1} + (1-\lambda)x^k) \le f(x^k) < f(p) + \sigma \quad \text{for all } \lambda \in [0,1]\,, \tag{4.5}$$

so $x^{k+1} \notin V$ would imply $\|x^{k+1} - p\| \ge \varepsilon$ and hence $\|\lambda x^{k+1} + (1-\lambda)x^k - p\| = \varepsilon$ for some $\lambda \in (0,1]$, contradicting the definition of σ. By compactness, all accumulation points of (x^k) must lie in B and thus in U. Since all of them are stationary by assumption, there can only be one, namely p, which means that (b) holds.

(b) ⇒ (a): Choose $U := V$. By monotonicity and continuity, we have $f(p) = \inf_k f(x^k) < f(x^0)$ for all $x^0 \in U \setminus \{p\}$, a set which does not contain any stationary point, as all sequences starting there have to converge to p by assumption. □

We thus have shown that in our model, every strict local solution is isolated (among all alternative stationary points $\tilde{x} \in \Delta$), which generally is not the case. Inspection of the proof of Theorem 4.3 reveals that the only essential property is that the iteration is *homotopical*, i.e. that the inequality on the left-hand side in (4.5) holds. We can conclude that for all these homotopical iteration procedures, convergence to a saddle point is highly unlikely, which is in line with recent findings in this research area for other first-order methods, see, e.g. [17] and references therein. Note that most of these papers deal with smooth transition maps (which facilitate characterization of saddle points via the Jacobian matrix) while our transition maps lack even continuity.

Next we need an auxiliary observation which only applies to d_{AFW}:

LEMMA 4.4. *Let Assumption 2.1 hold. Let $\gamma = \lim_{k\to\infty} f(x^k) = \inf_{k\in\mathbb{N}} f(x^k)$ and assume $\gamma = f(e^i)$ for some $i \in I$. Consider a certain iteration counter k with $x^{k+1} \ne x^k$. Then the following implications hold for both stepsize choices (3.6) or (3.8):*

(a) *if $d^k = d_{FW}^{x^k} = e^i - x^k$, then Algorithm 3.1 stops at $k+1$;*
(b) *if $x_i^k > 0$ and $d^k = \frac{x_i^k}{1-x_i^k}\, d_A^{x^k} = \frac{x_i^k}{1-x_i^k}\,(x^k - e^i)$, then $x_i^{k+1} = 0$.*

Proof. (a) By construction and assumption, we have $0 \le f(x^{k+1}) - \gamma \le f(e^i) - \gamma = 0$, hence $x^{k+2} = x^{k+1}$ which is a stationary point, using Proposition 3.1(a) or Corollary 3.2(a).

(b) Suppose that $\alpha_k < 1$; then φ_k has to be strictly convex, and by smoothness, f has to be strictly convex over the whole interval $conv(x^{k+1}, e^i)$. But as assumed above, we have $f(e^i) = \gamma \le f(x^{k+1}) < f(x^k)$ in contradiction to the fact that $x^k \in conv(x^{k+1}, e^i)$. So necessarily $\alpha_k = 1$ and therefore $x_i^{k+1} = 0$. □

We proceed to establish a convergence result for the full sequence of iterates under mild assumptions for the away-step Frank-Wolfe variant:

THEOREM 4.5. *Let Assumption 2.1 hold. Consider a sequence (x^k) generated by Algorithm 3.1 with stepsize choice (3.6) or (3.8), and d_{AFW} as descent direction. Suppose that (x^k) has finitely many accumulation points. Then it must converge: there is a $p \in \Delta$ such that $x^k \to p$ as $k \to \infty$.*

Proof. The statement obviously needs a proof only if the sequence (x^k) is infinite. So suppose there are finitely many (pairs of) accumulation points, but at least two. Choose pairwise disjoint neighbourhoods around all of them and wait until all x^k lie exactly in one of these neighbourhoods if $k \geq k_0$. Then, arguing by contradiction, if x^k would not converge, there is a subsequence k_j with $k_1 \geq k_0$ such that $x^{k_j} \to p$ and the immediate successors $x^{k_j+1} \to q \neq p$ as $j \to \infty$ which implies $\bar{\alpha} := \inf_j \alpha_{k_j} > 0$. Now, by thinning (k_j) if necessary, we may and do assume that $\alpha_{k_j} \to \alpha_\infty > 0$ as $j \to \infty$, and that there is an $i \in I$ with $d^{k_j} = e^i - x^{k_j}$ for all j, or else $d^{k_j} = \frac{x_i^{k_j}}{1-x_i^{k_j}} (x^{k_j} - e^i)$ with $x_i^{k_j} > 0$ for all j. Moreover, in this case we even get $x_i^{k_j} > c$ for all j and a suitable constant $c > 0$ because of

$$0 < \|q - p\| = \lim_j \|x^{k_j+1} - x^{k_j}\| \leq \sqrt{2}\alpha_\infty \lim_j \frac{x_i^{k_j}}{1 - x_i^{k_j}} = \sqrt{2}\alpha_\infty \frac{p_i}{1 - p_i} .$$

Next suppose that eventually stepsize is smaller than one, and we are in the strictly convex case. Then, employing (3.17) and

$$f(x^{k+1}) - f(x^k) = \varphi_k(\alpha_k) - \varphi_k(0) = \alpha_k \left[\dot{\varphi}_k(0) + \frac{\alpha_k}{2} \ddot{\varphi}_k(\hat{\alpha}_k) \right], \tag{4.6}$$

we obtain

$$\ddot{\varphi}_{k_j}(\hat{\alpha}_{k_j}) \to 0 . \tag{4.7}$$

Furthermore by continuity we have for any $\alpha \in I_{\text{feas}}(p, e^i - p)$,

$$\ddot{\varphi}_{k_j}(\alpha) \to (e^i - p)^\top \nabla^2 f((1-\alpha)p + \alpha e^i)(e^i - p),$$

in the FW case, and for any $\alpha \in I_{\text{feas}}(p, \mu(p - e^i))$,

$$\ddot{\varphi}_{k_j}(\alpha) \to \mu^2 (p - e^i)^\top \nabla^2 f((1+\alpha\mu)p - \alpha\mu e^i)(p - e^i),$$

with $\mu = \frac{p_i}{1-p_i}$, in the away step case. On the other hand, we can employ (4.7) in all cases. Now by (2.2) in Assumption 2.1 for $x = (1-\alpha)p + \alpha e^i, \alpha \in [0,1]$, and by choosing $d = e^i - p$ in the FW case or $d = \mu(p - e^i)$ in the away step case, we conclude that f must be linear along $conv(p, e^i)$ with slope $\lim_j \dot{\varphi}_{k_j}(0) = 0$, so constant, and $f(e^i) = f(p) = \inf_k f(x^k)$ results.

Now in case of the FW direction, Lemma 4.4(a) would then yield the absurd conclusion that Algorithm 3.1 stops even at iteration $k_1 + 1$.

In case of the away direction, we conclude by Lemma 4.4(b) that $x_i^{k_j+1} = 0$. But since $x_i^{k_{j+1}} > 0$, we must have a FW step $d^k = e^i - x^k$ for some $k \in \{k_j + 1, \ldots, k_{j+1} - 1\}$. Now we again invoke Lemma 4.4(a) to arrive at the contradiction that Algorithm 3.1 stops at iteration $k + 1$, using $f(e^i) = f(p) = \inf_k f(x^k)$.

So we are left with the case that the stepsize equals eventually one. But then the argument is even simpler: in the FW case, we stop at e^i, and in the away case we directly get $x_i^{k_j+1} = 0$ and, as argued just before, stop again at e^i at some iteration counter $k \in \{k_j + 1, \dots, k_{j+1} - 1\}$ as well. ∎

As a corollary to Theorems 4.1, 4.2 and 4.5, we thus obtain a generic convergence result for the iterates generated by Algorithm 3.1 for the away-step variant:

COROLLARY 4.6. *Suppose that (2.1) admits only finitely many stationary points. Then any sequence (x^k) generated by Algorithm 3.1 with stepsize choice (3.6) or (3.8), and d_{AFW} as descent direction, must converge: there is a $p \in \Delta$ such that $x^k \to p$ as $k \to \infty$.*

Now we introduce three sets that will be useful when carrying out the analysis related to support identification in finite time. More specifically, we call

$$S_+(x) := \{i \in I |\ \nabla_i f(x) > x^\top \nabla f(x)\},$$

$$S_-(x) := \{i \in I |\ \nabla_i f(x) < x^\top \nabla f(x)\},$$

and

$$S_0(x) := \{i \in I |\ \nabla_i f(x) = x^\top \nabla f(x)\}.$$

We hence report the announced results on support identification in finite time; note that strict complementarity (again generically true) of the stationary point $\bar{x}$ exactly means $\bar{S} = S_0(\bar{x})$ in below theorem; recall that $S_-(\bar{x}) = \emptyset$ by (3.1).

THEOREM 4.7. *Consider a convergent sequence of iterates (x^k), with supports $S_k = S(x^k)$, generated by Algorithm 3.1 for the following specifications:*

- *the search direction d^k is computed according to (AFW) or (PFW) rule;*
- *the stepsize α_k is computed using the line search described in (3.6) or (3.8).*

Denote by $\bar{x} := \lim_{k\to\infty} x^k$ as well as $\bar{S} := \{i \in I : \bar{x}_i > 0\}$, so that by stationarity (3.1) of $\bar{x}$ we have $\bar{S} \subseteq S_0(\bar{x})$. Then there is a finite $\bar{k}$ such that

$$\bar{S} \subseteq S_k \subseteq S_0(\bar{x}) \quad \text{for all } k \geq \bar{k}\,.$$

Proof. We can assume that $x^k = e^i$, with $i \in I$, cannot happen infinitely often. Indeed, otherwise by Lemma 4.4 the algorithm would stop after a finite number of iterations. So, we assume that $x^k \neq e^i$ for k sufficiently large. Now, by continuity of the gradient, we can find an iterate such that both the following inclusions hold:

$$S_+(\bar{x}) \subseteq S_+(x^k) \quad \text{and} \quad \bar{S} \subseteq S_k\,.$$

From stationarity of $\bar{x}$ we can further write $\bar{S} \subseteq S_0(\bar{x}) = I \setminus S_+(\bar{x})$. Hence, we have

$$S_-(x^k) \subseteq I \setminus S_+(x^k) \subseteq I \setminus S_+(\bar{x}) = S_0(\bar{x})$$

implying

$$S_-(x^k) \subseteq S_0(\bar{x})\,. \tag{4.8}$$

We claim now that once

$$S_k \subseteq S_0(\bar{x}) \tag{4.9}$$

holds for some k, then (4.9) is guaranteed for all the following iterations. Indeed, either $S_{k+1} = S_k \cup \{i\}$ and $i \in S_-(x^k) \subseteq S_0(\bar{x})$ or else the support is a subset of the current support, i.e., $S_{k+1} \subseteq S_k$. By contradiction to (4.9), let us assume that, when k sufficiently large, the set $S_k \setminus S_0(\bar{x})$ is never empty. Again, by continuity of the gradient, we can choose a sufficiently large k_0 to ensure existence of a positive value $\varrho > 0$ such that

$$|\nabla f(x^{k_j})^\top (e^i - x^{k_j})| < \varrho \quad \text{for all} \quad i \in S_0(\bar{x}) \quad \text{whenever} \quad k \geq k_0\,;$$

and

$$\nabla f(x^k)^\top (e^r - x^k) > \varrho \quad \text{for all } r \in S_k \setminus S_0(\bar{x}) = S_k \cap S_+(\bar{x}) \quad \text{whenever} \quad k \geq k_0\,.$$

Hence, for both direction variants (AFW) and (PFW), we have that $e^{r(k)}$ is chosen in the algorithm as away-step vertex for some $r(k) \in S_k \setminus S_0(\bar{x})$, if $k \geq k_0$. Further, due to the finiteness of I, by considering a suitable subsequence k_j we can assume

$$r(k_j) = r \in S_{k_j} \setminus S_0(\bar{x}) = S_{k_j} \cap S_+(\bar{x})\,.$$

By stationarity of $\bar{x}$ we know $r \notin \bar{S}$, so the r-th coordinate of $\bar{x}$ satisfies $\bar{x}_r = 0$. Eventually, $x_r^{k_j+1} = 0$ holds exactly because otherwise $\alpha_{k_j} < 1$ and Prop. A.1 or Prop. A.2 applies, contradicting $x_r^{k_j} \to \bar{x}_r = 0$. Repeating the same argument for all other indices in $S_k \setminus S_0(\bar{x})$, the result is proved. ∎

As we pointed out in the introduction, the classic Frank-Wolfe algorithm does not guarantee support identification in finite time. Below we report an example where all iterates x^k have full support (i.e., $|S^k| = n$) and the point $\bar{x}$ does not (i.e., $|\bar{S}| < n$).

EXAMPLE 4.1 (Bad behaviour of the Frank-Wolfe algorithm). *We consider problem (2.1) having a quadratic objective function*

$$f(x) = \frac{1}{2} x^\top Q x\,,$$

with

$$Q = \begin{bmatrix} 6 & 0 & 6 \\ 0 & 3 & 3 \\ 6 & 3 & 10 \end{bmatrix}.$$

It is easy to see that the solution in this case is the global minimizer $\bar{x} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix}^\top$. *If we choose as starting point* $x^0 = (0.1\ 0.3\ 0.6)^\top$, *the Frank-Wolfe algorithm will not be able to get an iterate with the same support as* $\bar{x}$ *in finite time, neither via exact nor with Armijo line search [8].*

Moreover, the behaviour of this version may be even deceptive as the support of the iterates is eventually constant also for this algorithm; indeed, either the iterates coincide with a vertex e^i infinitely often, so that monotonicity would imply finite convergence to e^i. But this is the benevolent case. In the opposite case, eventually no vertex is hit exactly during iterations, so that supports must (weakly) increase with t. By finiteness it follows that they remain eventually constant, but, as the example shows, S^k may overestimate the correct support $\bar{S}$.

5. Conclusions. In this paper, we studied methods for solving minimization problems over the probability simplex. More specifically, we analyzed two variants of the Frank-Wolfe algorithm, namely away-step and pairwise Frank-Wolfe. We first proved convergence of the iterates to stationary points both when using exact and Armijo line search, and even convergence for the full sequence of iterates for the away-step variant, under mild regularity assumptions. Then we showed that both discussed variants algorithms guarantee support identification in finite time, a property shared by projected gradient methods. As a future development, it may be worth while to analyze conditions which allow to get explicit bounds on the number of iterations required to identify the support correctly.

Appendix A. Auxiliary results.

PROPOSITION A.1. *Let $(x^{s_j}) \to x^*$ as $j \to \infty$ be a convergent subsequence generated by Algorithm 3.1 according to (AFW) or (PFW) rule, where we abbreviate $d^j_{FW} = d^{x^{s_j}}_{FW}$ and $d^j_A = d^{x^{s_j}}_A$. We assume that for some fixed r, we have for all j*

- $d^{x^{s_j}}_{AFW} = \frac{x^{s_j}_r}{1-x^{s_j}_r}\, d^j_A = \frac{x^{s_j}_r}{1-x^{s_j}_r}\,(x^{s_j} - e^r)$ *or* $d^{x^{s_j}}_{PFW} = x^{s_j}_r (e^{i_j} - e^r)$;
- *the stepsize is computed using the line search described in (3.6) and satisfies* $\alpha_{s_j} < 1$;
- *one of the following cases holds:*
  1. *there exists i such that $\nabla_i f(x^*) < \nabla f(x^*)^\top x^*$, or*
  2. *there exists $\varrho > 0$ such that $\nabla f(x^{s_j})^\top (e^r - x^{s_j}) > \varrho$.*

Then $x^*_r > 0$.

Proof. Since $\alpha_{s_j} < 1$ we have that (3.5) holds for some $\tilde\alpha_{k_j} \in [0,1]$. So we arrive via (2.3) at

$$\frac{x^{s_j}_r}{1-x^{s_j}_r} \geq \alpha_{s_j} \frac{x^{s_j}_r}{1-x^{s_j}_r} = \frac{-\dot\varphi_{s_j}(0)}{\ddot\varphi_{s_j}(\tilde\alpha_{s_j})} \frac{x^{s_j}_r}{1-x^{s_j}_r} = \frac{-\nabla f(x^{s_j})^\top d^j_A}{[d^j_A]^\top \nabla^2 f(x^{s_j} + \tilde\alpha_{k_j} d^{s_j}) d^j_A}$$

$$\geq \frac{-\nabla f(x^{s_j})^\top d^j_A}{2K} \geq \frac{-\nabla f(x^{s_j})^\top d^j_{FW}}{2K} \geq \frac{\nabla f(x^{s_j})^\top x^{s_j} - \nabla_i f(x^{s_j})}{2K}$$

$$\to \frac{\nabla f(x^*)^\top x^* - \nabla_i f(x^*)}{2K} > 0 \quad \text{as } j \to \infty\,,$$

if we assume that Case 1 holds, and we get the same inequality for $j \to \infty$ also in Case 2 since

$$-\nabla f(x^{s_j})^\top d^j_A > \varrho > 0\,.$$

This implies $x^*_r > 0$ for the (AFW) rule and likewise

$$x^{s_j}_r \geq \alpha_{s_j} x^{s_j}_r = \frac{-\dot\varphi_{s_j}(0)}{\ddot\varphi_{s_j}(\tilde\alpha_{s_j})}\, x^{s_j}_r = \frac{-\nabla f(x^{s_j})^\top [d^j_{FW} + d^j_A]}{[d^j_{FW} + d^j_A]^\top \nabla^2 f(x^{s_j} + \tilde\alpha_{k_j} d^{s_j})[d^j_{FW} + d^j_A]}$$

$$\geq \frac{-\nabla f(x^{s_j})^\top [d^j_{FW} + d^j_A]}{2K} \geq \frac{-\nabla f(x^{s_j})^\top d^j_{FW}}{2K} \geq \frac{\nabla f(x^{s_j})^\top x^{s_j} - \nabla_i f(x^{s_j})}{2K}$$

$$\to \frac{\nabla f(x^*)^\top x^* - \nabla_i f(x^*)}{2K} > 0 \quad \text{as } j \to \infty$$

proves the result in Case 1 with the (PFW) rule, and the same inequality for $j \to \infty$ holds in Case 2 since

$$-\nabla f(x^{s_j})^\top [d_{FW}^j + d_A^j] > -\nabla f(x^{s_j})^\top d_A^j > \varrho > 0\,. \qquad \square$$

PROPOSITION A.2. *Let $(x^{s_j}) \to x^*$ as $j \to \infty$ be a convergent subsequence generated by Algorithm 3.1 according to (AFW) or (PFW) rule, where we abbreviate $d_{FW}^j = d_{FW}^{x^{s_j}}$ and $d_A^j = d_A^{x^{s_j}}$. We assume that for some fixed r, we have for all j*

- $d_{AFW}^{x^{s_j}} = \frac{x_r^{s_j}}{1-x_r^{s_j}}\, d_A^j = \frac{x_r^{s_j}}{1-x_r^{s_j}}\,(x^{s_j} - e^r)$ *or* $d_{PFW}^{x^{s_j}} = x_r^{s_j}(e^{i_j} - e^r)$;
- *the stepsize is computed using the Armijo line search described in (3.8) and satisfies $\alpha_{s_j} < 1$;*
- *one of the following cases holds:*
  1. *there exists i such that $\nabla_i f(x^*) < \nabla f(x^*)^\top x^*$, or*
  2. *there exists $\varrho > 0$ such that $\nabla f(x^{s_j})^\top (e^r - x^{s_j}) > \varrho$.*

Then $x_r^* > 0$.

Proof. We first notice that for any $\alpha \in [0,1]$ and $k = s_j$, by (2.3) we can write

$$f(x^k + \alpha d^k) \le f(x^k) + \alpha^k \nabla f(x^k)^\top d^k + \frac{\alpha^2 K}{2}\|d^k\|^2\,.$$

So the sufficient decrease condition (3.7) would be satisfied if

$$f(x^k) + \alpha \nabla f(x^k)^\top d^k + \frac{\alpha^2 K}{2}\|d^k\|^2 \le f(x^k) + \gamma\alpha\nabla f(x^k)^\top d^k\,,$$

and the latter holds true if

$$\alpha \le \alpha_k^{\max} := \frac{2(1-\gamma)}{K}\frac{|\nabla f(x^k)^\top d^k|}{\|d^k\|^2}.$$

This gives us an interval $[0, \alpha_k^{\max}]$ of step sizes satisfying sufficient decrease. Now, if $\alpha_k < 1$ is chosen as Armijo step size, then either $\alpha_k > \alpha_k^{\max}$ or else $\alpha_k \in [0, \alpha_k^{\max}]$ but then $\frac{\alpha_k}{\delta} > \alpha_k^{\max}$ as the step size $\alpha = \frac{\alpha_k}{\delta}$ shall violate (3.7) by definition (3.8). In both cases, we get

$$\alpha_k > \delta\alpha_k^{\max}\,.$$

Now we consider the two different search directions, abbreviating $d_{FW}^j = d_{FW}^{x^{s_j}}$ and

$d_A^j = d_A^{x^{s_j}}$. For the (AFW) rule, Case 1, we can write

$$\frac{x_r^{s_j}}{1-x_r^{s_j}} \geq \alpha_{s_j} \frac{x_r^{s_j}}{1-x_r^{s_j}} > \delta\, \alpha_{s_j}^{\max} \frac{x_r^{s_j}}{1-x_r^{s_j}}$$

$$= \frac{-\nabla f(x^{s_j})^\top d_A^j}{\|d_A^j\|^2} \frac{2\delta(1-\gamma)}{K} \geq \frac{-\nabla f(x^{s_j})^\top d_A^j}{2} \frac{2\delta(1-\gamma)}{K}$$

$$\geq \frac{-\nabla f(x^{s_j})^\top d_{FW}^j}{2} \frac{2\delta(1-\gamma)}{K}$$

$$\geq \frac{\delta(1-\gamma)}{K} \left[\nabla f(x^{s_j})^\top x^{s_j} - \nabla_i f(x^{s_j})\right]$$

$$\to \frac{\delta(1-\gamma)}{K} \left[\nabla f(x^*)^\top x^* - \nabla_i f(x^*)\right] > 0 \text{ as } j \to \infty\,,$$

and the same inequality for $j \to \infty$ can be obtained in Case 2 since

$$-\nabla f(x^{s_j})^\top d_A^j > \varrho\,.$$

This implies $x_r^* > 0$. Similarly, for (PFW), Case 1, we can write

$$x_r^{s_j} \geq \alpha_{s_j} x_r^{s_j} > \delta\, \alpha_{s_j}^{\max} x_r^{s_j}$$

$$= \frac{-\nabla f(x^{s_j})^\top [d_A^j + d_{FW}^j]}{\|d_A^j + d_{FW}^j\|^2} \frac{2\delta(1-\gamma)}{K} \geq -\nabla f(x^{s_j})^\top [d_A^j + d_{FW}^j] \frac{\delta(1-\gamma)}{K}$$

$$\geq -\nabla f(x^{s_j})^\top d_{FW}^{x^{s_j}} \frac{\delta(1-\gamma)}{K}$$

$$\geq \frac{\delta(1-\gamma)}{K} \left[\nabla f(x^{s_j})^\top x^{s_j} - \nabla_i f(x^{s_j})\right]$$

$$\to \frac{\delta(1-\gamma)}{K} \left[\nabla f(x^*)^\top x^* - \nabla_i f(x^*)\right] > 0 \text{ as } j \to \infty\,,$$

and the same inequality holds for $j \to \infty$ in Case 2 since

$$-\nabla f(x^{s_j})^\top [d_{FW}^j + d_A^j] > -\nabla f(x^{s_j})^\top d_A^j > \varrho > 0\,. \qquad \square$$

REFERENCES

[1] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Transactions on automatic control, 21 (1976), pp. 174–184.

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

[3] D. P. Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.

[4] 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.

[5] J. Burke, *On the identification of active constraints ii: The nonconvex case*, SIAM Journal on Numerical Analysis, 27 (1990), pp. 1081–1102.
[6] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM Journal on Numerical Analysis, 25 (1988), pp. 1197–1211.
[7] K. L. Clarkson, *Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm*, ACM Transactions on Algorithms (TALG), 6 (2010), p. 63.
[8] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, *New active-set frank-wolfe variants for minimization over the simplex and the ℓ_1-ball*, arXiv preprint arXiv:1703.07761, (2017).
[9] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, *A two-stage active-set algorithm for bound-constrained optimization*, J. Optim. Theory Appl., 172 (2017), pp. 369–401.
[10] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, *An active-set algorithmic framework for non-convex optimization problems over the simplex*, arXiv preprint arXiv:1703.07761v2, (2018).
[11] E. de Klerk, *The complexity of optimizing over a simplex, hypercube or sphere: a short survey*, Central European Journal of Operations Research, 16 (2008), pp. 111–125.
[12] M. De Santis, G. Di Pillo, and S. Lucidi, *An active set feasible method for large-scale minimization problems with bound constraints*, Computational Optimization and Applications, 53 (2012), pp. 395–423.
[13] J. Guélat and P. Marcotte, *Some comments on Wolfe's away step*, Mathematical Programming, 35 (1986), pp. 110–119.
[14] W. W. Hager and H. Zhang, *A new active set algorithm for box constrained optimization*, SIAM J. Optim., 17 (2006), pp. 526–557.
[15] W. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.
[16] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of Frank-Wolfe optimization variants*, in NIPS 2015 - Advances in Neural Information Processing Systems, 2015.
[17] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid saddle points*, arXiv preprint arXiv:1710.07406, (2017).
[18] S. Lee and S. J. Wright, *Manifold identification in dual averaging for regularized stochastic online learning*, Journal of Machine Learning Research, 13 (2012), pp. 1705–1744.
[19] R. Mifflin and C. Sagastizábal, *Proximal points are on the fast track*, Journal of Convex Analysis, 9 (2002), pp. 563–580.
[20] B. Mitchell, V. F. Dem'yanov, and V. Malozemov, *Finding the point of a polyhedron closest to the origin*, SIAM Journal on Control, 12 (1974), pp. 19–26.
[21] J. Nutini, M. Schmidt, and W. Hare, *"Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?*, Optimization Letters, (2018), https://doi.org/10.1007/s11590-018-1325-z.
[22] J. She and M. Schmidt, *Linear convergence and support vector identification of sequential minimal optimization*, 10th NIPS Workshop on Optimization for Machine Learning, (2017).
[23] P. Wolfe, *Convergence theory in nonlinear programming*, Integer and nonlinear programming, (1970), pp. 1–36.
[24] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1063–1079.
[25] S. J. Wright, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM Journal on Optimization, 22 (2012), pp. 159–186.