University of Padua

Department of Mathematics "Tullio Levi-Civita"

Master's degree in mathematics

Convergence analysis and active set complexity for some FW variants

Supervisor: Francesco Rinaldi

Author: Damiano Zeffiro **Matricola No.**: 1152132

July 4, 2019

Contents

Chapter 1

Introduction

1.1 Related work

Consider the smooth convex optimization problem

$$\min\{f(x) \mid x \in \Omega\} \tag{1.1.1}$$

where Ω is a compact and convex subset of $\mathbb{R}^n$, f is a continuously differentiable convex function, and we define $X^* \subset \Omega$ as the set of solutions.
Among many first order algorithms to solve this problem the FW algorithm, first introduced in 1956 by Marguerite Frank and Philip Wolfe [21], has recently been the subject of renewed interest.
There are two main properties of this algorithm that make it more suitable for many smooth convex optimization problems than other first order optimization methods. Let $\{x_k\}_{k\in\mathbb{N}}$ be the sequence generated by the FW method. The first property is that x_k is the convex combination of at most k "elementary" points, and for many problems this means that the FW method can find efficiently e.g. low rank and sparse solutions. The second property is that the main computational cost of each iteration comes from solving a problem with linear objective to compute the search direction, that is

$$d_k^{FW} = y - x_k \text{ with } y \in \operatorname{argmin}\{(\nabla f(x_k), y) \mid y \in \Omega\} \tag{1.1.2}$$

When Ω is a polytope computing d_k^{FW} becomes a LP problem, which is often cheaper than projecting on Ω. FW like algorithms have been successfully applied to optimization over polyhedral sets in submodular function optimization [1], structured SVM learning [35] and variational inference problems [31]. For these problems the feasible region is represented as the convex hull of a finite set of points. However, FW like algorithms have also been applied to smooth optimization over non polyhedral sets for instance in matrix completion [22] and metric learning [14] problems.
The main thread of the thesis is proving convergence properties for some popular variations of the classic FW method. The theoretical analysis of FW like algorithms is currently a very active area of research, where much work has been done to prove that FW variants can match the convergence rate of other first order methods. In this thesis we prove not only general convergence rates for the value of the objective function, but also results about active set complexity.
We now discuss the convergence results for FW like algorithms most relevant to the

thesis, while we refer the reader to [23] for a more detailed summary.
One of the drawbacks of the classic FW algorithm is its slow theoretical convergence rate on convex compact sets of $O(1/k)$ [21], which can be tight even for polytopes and quadratic objective functions when the solution is on the boundary [13].
The slow convergence rate of the FW algorithm when the solutions are a subset of the boundary is due to the fact that as the algorithm approaches the boundary the search directions can become almost orthogonal to the gradient, and the sequence $\{x_k\}_{k\in\mathbb{N}}$ then forms a slow converging zig-zag pattern. To address this problem Wolfe [42] formulated the away step FW (AFW), which at every step choses between the classic FW direction given by (2.5.68) and an alternative search direction

$$d_k^{AW} = x_k - y \text{ with } y \in \operatorname{argmax}\{(\nabla f(x_k), u) \mid u \in S_k\} \tag{1.1.3}$$

where $S_k \subset \Omega$ is such that $x_k \in \operatorname{conv}(S_k)$ and $|S_k| \leq k$. It was proved only recently ([33], [34]) that the AFW has a linear convergence rate on polytopes for μ - strongly convex objectives with $L-$ Lipschitz differential. This rate depends from the condition number of the objective $\frac{\mu}{L}$ and from a certain parameter $\mathrm{PdirW}(\Omega)$ determined by the polytope. This result was then extended to other FW variants and also for objective of the form $f(x) = g(Ax) + (b, x)$ with g strongly convex [3]. Later [39] several equivalent characterization for $\mathrm{PdirW}(\Omega)$ were proved.
As for non polyhedral sets, a convergence rate of $O(\frac{1}{k^2})$ was proved for the classic FW method applied to strongly convex functions on strongly convex sets [23]. In order to interpolate between the convergence rate in the general convex case and the convergence rate of the strongly convex one the Holderian error bound condition

$$f(x) - f^* \geq \gamma \operatorname{dist}(x, X^*)^p \tag{1.1.4}$$

was used in the recent works [44], [30] where $\gamma > 0$, $p \geq 1$ and f^* is the minimum of $f|_\Omega$. Finally, for non convex smooth functions a convergence rate of $O(\frac{1}{\sqrt{k}})$ was proved [32] for the FW gap

$$g^{FW}(x_k) = (\nabla f(x_k), x_k) - \min\{y \in \Omega \mid (\nabla f(x_k), y)\} \tag{1.1.5}$$

The active set identification problem (AS identification problem) is broadly speaking identifying the manifold containing the set of minimizers or more in general a certain subset of stationary points. In this thesis we focus on a geometric definition for the AS identification problem. Given $\bar{x} \in X^*$ we say that the manifold containing $\bar{x}$ is the face $E_\Omega(-\nabla f(\bar{x}))$ of Ω exposed by $-\nabla f(x)$:

$$E_\Omega(-\nabla f(\bar{x})) = \{x \in \Omega \mid (-\nabla f(\bar{x}), x - \bar{x}) = 0\} \tag{1.1.6}$$

The set $E_\Omega(-\nabla f(\bar{x}))$ is a face of Ω since by first order optimality conditions

$$(-\nabla f(\bar{x}), x - \bar{x}) \leq 0 \ \forall \ x \in \Omega \tag{1.1.7}$$

When f is convex it is not difficult to see that $E_\Omega(-\nabla f(x))$ does not depend on the particular $x \in X^*$, so that we can define the support of X^* as $A_f^\Omega = E_\Omega(-\nabla f(x))$ for any $x \in X^*$. A sequence generated by a certain method identifies a support if the

sequence generated by the method is definitely in the support.
The projected gradient algorithm applied to convex functions on polytopes identifies A^f_Ω in finite time ([11], [12]). An analogous result was recently proved for the AFW and the PFW [9] using hypotheses on f weaker than strong convexity but which are not implied by convexity alone (and conversely). Another possible approach to identify a support is to combine a first order method with an AS strategy which can identify the support of a solution once the sequence is close enough to that solution, improving the value of the objective at the same time. Such a strategy was recently defined on the simplex [18] and can be generalized to polytopes when combined with an affine invariant method like the FW or the AFW methods.

1.2 Contributions

We now describe the original contributions of this thesis chapter by chapter.

Chapter 2. We analyze the FW method with in face directions (FDFW), which was originally introduced for polyhedral sets [27], but which we apply to general convex sets. A variation of this method was recently applied to matrix completion problems, with numerical results showing that it outperforms other FW variants [22]. For every k the algorithm choses its search direction between the FW one and another feasible for the minimal face of Ω containing x_k. Our main reason for choosing this method is that the AFW linear convergence rate property does not seem to extend to non polyhedral sets, as observed by Lacoste-Julien and Jaggi in their article proving linear convergence for the AFW on polytopes [34]. Instead we prove that the FDFW has a linear convergence rate not only on polytopes, but also on a class of strictly convex sets. As for the AFW, this linear convergence rate depends on the condition number of f and on a parameter determined by the geometry of the feasible set Ω, which we call $\mathrm{NW}(\Omega)$. In particular when f is strongly convex with Lipschitz gradient we have linear convergence whenever $\mathrm{NW}(\Omega) > 0$. The idea behind the proof is basically that $\mathrm{NW}(\Omega)$ is designed to ensure that the slope along the direction computed by the FDFW is at least a fraction depending on $\mathrm{NW}(\Omega)$ of the highest slope possible among all feasible descent directions. These properties of slopes then translate to a linear convergence rate because $f(x_k)-f(x^*)$ and $f(x_k)-f(x_{k+1})$ can be upper and lower bounded proportionally to the slope of $-\nabla f(x_k)$ along $x^* - x_k$ and $x_{k+1} - x_k$ respectively for $x^* \in X^*$. We prove that for polytopes $\mathrm{NW}(\Omega) \geq \frac{\mathrm{PdirW}(\Omega)}{2D}$ with $D = \mathrm{diam}(\Omega)$ applying the alternative characterization of $\mathrm{PdirW}(\Omega)$ recently proved in [39]. Then we prove $\mathrm{NW}(\Omega) > 0$ also for some strictly convex sets including those whose boundary looks locally like a sphere. We do this by giving a lower bound on $\mathrm{NW}(\Omega)$ with an expression involving the variation of

$$l_{\Omega,\bar{x}}(d) = \max\{\lambda \in R_{\geq 0} \mid \bar{x} + \lambda d \in \Omega\} \tag{1.2.1}$$

for a fixed $\bar{x} \in \partial\Omega$ as a function of $\mathrm{dist}(d, T_\Omega(\bar{x})^c)$, where $T_\Omega(\bar{x})$ is the tangent cone to Ω in the point $\bar{x}$ and d can vary in $\partial B(0,1)$.
In the convergence analysis we assume the Holderian error bound condition (1.1.4) to interpolate between the $O(\frac{1}{k})$ and the $O(e^{-\lambda k})$ rates of the general

and strongly convex case respectively.
The main drawback of the FDFW is that even its "simplest" variant needs oracles for maximal feasible step sizes and linear optimization in the minimal face containing the current iterate. These oracles may be very expensive when Ω is represented as the convex hull of a finite number of points. In this setting we show how at least in theory computing these oracles can be reduced to other well known optimization problems.

Chapter 3. We first prove the equivalence between the geometric definition of support introduced in section 1 and an algebraic one based on non zero Lagrangian multipliers in some KKT like optimality conditions. Then we prove several AS related results for first order algorithms. First, we give a new proof for the projected gradient method finite time AS identification with explicit estimates using a property of polyhedral cones, that is the bijection between faces of a cone and faces of the dual cone given by the orthogonality mapping. Our proof however still needs the Moreau-Yosida lemma used in the original proof. We then prove finite time AS identification for non convex objectives for the AFW and the PFW methods assuming convergence to a subset of X^* with constant support.

Chapter 4. In this chapter we give explicit estimates for the AS identification complexity of the AFW on polytopes. We work considering as feasible region the simplex and then prove analogous results for generic polytopes mainly using the affine invariance of the AFW. With respect to the recent work done on AFW complexity [9] our main improvement is removing the additional assumptions on f, which in our work can be any function with Lipschitz gradient. We also give explicit bounds for the AS radius as a function of the Lipschitz constant of $\nabla f(x)$ and the value of Lagrangian multipliers. This AS radius is the radius of a ball centered on a point $x^* \in X^*$ inside which the AFW identifies "quickly" the support of x^*. We give a lower bound for the AS radius by approximating the optimal value of a related linear programming problem. As in [9], we make an additional assumption on the set of accumulation points for $\{x_k\}_{k\in\mathbb{N}}$ which generalizes convergence to a strict minimum. Finally, we obtain a general AS complexity bound that we combine to AFW converge rates for both strongly convex and non convex objectives to prove more explicit AS complexity bounds in these particular settings.

Chapter 2

A FW like algorithm with linear convergence on non polyhedra

The classic Frank Wolfe algorithm has a theoretical convergence rate of $O(1/t)$ for convex functions on convex sets. This rate is known to be tight on polytopes. At the same time, many variants of the Frank Wolfe algorithm have recently been shown to have a global linear convergence property on polytopes for strongly convex objectives. However, there are not many results concerning FW-like algorithms on non polytopes that give faster convergence rates using only the condition number of the objective and the geometry of the set.
In this chapter we analyze an algorithm which at each step chooses between the classic FW direction and an alternative direction in the minimal face containing the current iterate. We prove that our algorithm has global linear convergence rate for strongly convex objectives not only on polytopes, but also on a class of convex sets including strictly convex sets whose boundary locally looks like a sphere.

2.1 Introduction

The main focus of this chapter is to study the convergence rate of the in face directions FW method (FDFW) not only on polytopes, but also on a class of strictly convex sets. Following the techniques already used in [34], [39] and [33] for polytopes, we prove a linear convergence rate for strongly convex objectives. This rate depends on a parameter which we call normalized width resembling the pyramidal width defined for the first time in [33]. In [23] a convergence rate of $O(1/t^2)$ was proved for the classic FW algorithm on strongly convex sets for strongly convex functions. Our results can be viewed as an improvement of the convergence rate for a FW like algorithm on a class of strictly convex sets.
In section 2.2 we describe the FDFW algorithm and analyze several way to choose the in face directions. In section 2.3 we recall a few key definitions and basic properties of Holderian error bounds. In section 2.4 we give bounds for the normalized width on polytopes and a class of strictly convex sets. In section 2.5 we compute the convergence rate of the FDFW as a function of the Holderian error bound on the objective. In the appendix we recall a few useful theorems and definitions concerning convex sets and generalize part of our analysis to reflexive Banach spaces.

2.1.1 Notation

In the rest of the thesis Ω will be a compact and convex set unless specified otherwise, $f \in C^1(\Omega)$ a convex function with $\min f|_\Omega = f^*$ and $\text{argmin} f|_\Omega = X$, $\hat{c} = c/\|c\|$ for $c \in \mathbb{R}^N/\{0\}$, $\hat{\cdot}$ will always denote a vector in the unit euclidean ball; for $\bar{x} \in \Omega$ we define $T_\Omega(\bar{x})$ as the tangent cone to Ω in $\bar{x}$ (see subsection 5.1 of the appendix for some useful properties), and we also define $\mathcal{F}(x)$ as the minimal face of Ω containing x. We will indicate with Δ_n the $n-$ dimensional simplex $\Delta_n = \{x \in \mathbb{R}^n_{\geq 0} \mid \|x\|_1 = 1\}$. Finally, for a convex set C and a vector $r \in \mathbb{R}^n$ we define $\pi(C, r)$ as the projection of r on C.

2.2 FW method with in face directions

The in face direction FW method (FDFW) was introduced for the first time in [27] for polytopes represented as $\Omega = \{x \in \mathbb{R}^n \setminus \{0\} \mid Ax = b, x \geq 0\}$. Thanks to a simple geometric interpretation relying on the concept of minimial face it can easily be extended to generic compact convex sets. It is a FW method in the sense that its most expensive component (at least on polytopes) is the linear minimization oracle

$$\text{LMO}_C(r) \in \text{argmin}_{x \in C}(r, x) \tag{2.2.1}$$

where C can vary among the faces of Ω.
However, it also needs a stepsize oracle

$$\alpha_{\max}(\bar{x}, d) = \max\{\alpha \in \mathbb{R} \mid \bar{x} + \alpha d \in \Omega\} \tag{2.2.2}$$

for $\bar{x} \in \Omega$, $d \in \mathbb{R}^n$. This oracle gives the maximal feasible step in the direction d from $\bar{x}$. Notice that if $\Omega = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ then

$$\alpha_{\max}(\bar{x}, d) = \min\{\frac{b_i - A_i \bar{x}}{A_i d} \mid A_i d > 0\} \tag{2.2.3}$$

The FDFW, with new strategies to select the in face direction, was recently applied in [22] to the matrix completion problem. Some variants of the algorithm performed significantly better than the FW and the away step FW methods on the test problems.

Table 1: FW method with in face directions
1. Let $x_0 \in \Omega$
2. **for** k = 0...T **do**
3. Let $s_k := \text{LMO}_\Omega(\nabla f(x_k))$ and $d_k^{FW} := s_k - x_k$
4. Select d_k^A such that $x_k + d_k^A \in \text{aff}(\mathcal{F}(x_k))$ and $(\nabla f(x_k), d_k^A) \leq 0$
5. **if** $g_k^{FW} := (-\nabla f(x_k), d_k^{FW}) \leq \epsilon$ **then return** x_k
6. Choose $d_k \in \{d_k^A, d_k^{FW}\}$ using a suitable criterion.
7. **if** $d_k = d_k^{FW}$ **then** $\alpha_{\max} := 1$
8. **else** $\alpha_{\max} := \max\{\alpha \in \mathbb{R} \mid x_k + \alpha d_k^A \in \Omega\} = \alpha_{\max}(x_k, d_k)$
9. **end if**
10. Choose $\alpha_k \in [0, \alpha_{\max}]$ using e.g. line search.
11. Update $x_{k+1} := x_k + \alpha_k d_k$
12. **end for**

In table 1 a general scheme for the FDFW is described. In steps 3 and 4 the classical FW directions and the facial direction are computed respectively. In step 5 we have the stopping criterion with ϵ representing the desired precision on the objective: indeed it follows immediately from the properties of convex functions that

$$g_k^{FW} = (-\nabla f(x_k), d_k^{FW}) \geq f(x_k) - f^* \tag{2.2.4}$$

In step 6 the algorithm chooses between the facial direction and the classic FW direction according to a suitable criterion, while in step 7 to 10 the algorithm defines the step size which of course must never be greater than the maximal feasible step.
It remains to specify how to select d_k^A and how to choose between d_k^{FW} and d_k^A. Broadly speaking, there are two possible reasons to look for an alternative direction d_k^A. The first is that the classical FW direction can be almost orthogonal to the gradient, so that a line search along the alternative direction d_k^A can guarantee a greater decrease of the objective function. The second reason is that since $x_k + d_k^A$ is in $\mathcal{F}(x_k)$ a maximal step along d_k^A always decreases the dimension of the minimal face containing the current iterate. This is particularly useful whenever the solution $\bar{x}$ lies in a low dimensional face, since in this setting a FDFW method can hopefully identify $\mathcal{F}(\bar{x})$ in a finite number of steps.
We now describe three ways to define d_k^A. The first is

$$d_k^A = x_k - x_A \text{ with } x_A \in \text{argmax}\{(\nabla f(x_k), x) \mid x \in \mathcal{F}(x_k)\} \tag{2.2.5}$$

This choice is strictly related to the away direction selected by the AFW method. Indeed, if we further impose that $d_k^A = q_k - x_k$ with q_k a vertex of $\mathcal{F}(x_k)$, then d_k^A is an away direction with respect to the active set of atoms formed by all the vertexes in $\mathcal{F}(x_k)$.
Finally, one can always select the steepest descent direction possible in $\mathcal{F}(x_k)$, which is

$$d_k^A = x_p - x_k \text{ with } x_p = \pi_{\text{aff}(\mathcal{F}(x_k))}(x_k - \nabla f(x_k)) \tag{2.2.6}$$

This can be convenient when it is possible to compute the projection on $\text{aff}(\mathcal{F}(x_k))$ quickly, for instance when $\mathcal{F}(x_k)$ has low codimension.
As for the criterion to choose between d_k^A and d_k^{FW}, one can either compare the slopes

$$\textbf{if } (-\nabla f(x_k), \frac{d_k^{FW}}{\|d_k^{FW}\|}) \geq (-\nabla f(x_k), \frac{d_k^A}{\|d_k^A\|}) \textbf{ then } d_k = d_k^{FW}, \textbf{ else } d_k = d_k^A \tag{2.2.7}$$

or even without normalizing

$$\textbf{if } (-\nabla f(x_k), d_k^{FW}) \geq (-\nabla f(x_k), d_k^A) \textbf{ then } d_k = d_k^{FW}, \textbf{ else } d_k = d_k^A \tag{2.2.8}$$

with no significant differences in the theoretical analysis. To try and decrease quickly the dimension of $\mathcal{F}(x_k)$ a more aggressive strategy can be

$$\textbf{if } f(x_k + \alpha_{\max}(x_k, d_k^A)d_k^A) \leq f(x_k) \textbf{ then } d_k = d_k^A,\ x_{k+1} = x_k + \alpha_{\max}(x_k, d_k^A)d_k^A \tag{2.2.9}$$

followed by (2.2.7) or (2.2.8) whenever the condition $f(x_k + \alpha_{\max}(x_k, d_k^A)d_k^A) \leq f(x_k)$ is not satisfied.
As for the step size, we need a strategy that guarantees either a maximal step or a

decrease of the objective function proportional to the square of the slope along the descent direction. To this aim we can either use linesearch or compute the step size with the formula

$$\alpha_k = \min(\frac{(-\nabla f(x_k), d_k)}{\|d_k\|^2 L}, \alpha_{\max}) \tag{2.2.10}$$

2.3 Error bounds

A very common hypothesis used in convex optimization to obtain faster convergence results is strong convexity. However, this hypothesis leaves a gap between the slow convergence rate of the general convex case and the fast convergence rate of the strongly convex one. In order to interpolate between these convergence rates one possibility is to use Holderian error bounds [8]. We now recall that for $y \in \mathbb{R}^n$ and $A \subset \mathbb{R}^n$ the point set distance $\mathrm{dist}(y, A)$ is defined as

$$\mathrm{dist}(y, A) = \inf\{\|y - z\| \mid z \in A\} \tag{2.3.1}$$

A continuous function $g : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy an error bound condition on the set Ω if it has minimum g^* with nonempty set of minimizers S and

$$w(g(x) - g^*) \geq \mathrm{dist}(x, S) \tag{2.3.2}$$

for every $x \in \Omega$ for some increasing $w : \mathbb{R} \to \mathbb{R}$ with $w(0) = 0$. When $w(x) = \beta x^\theta$ so that

$$\beta(g(x) - g^*)^\theta \geq \mathrm{dist}(x, S) \tag{2.3.3}$$

with $0 \leq \theta < 1$, $\beta > 0$ the error bound is said to be *Holderian.* When $\theta > 0$ this condition is sometimes written as

$$g(x) - g^* \geq \gamma \mathrm{dist}(x, S)^p \tag{2.3.4}$$

for some $\gamma > 0$ and $p = 1/\theta$. In the recent work [30] the Holderian error bound hypothesis was applied to the analysis of an AFW variant on polytopes. Convergence rates of $O(1/k^{\frac{1-\theta}{1-2\theta}})$ for $0 \leq \theta < 1/2$ and $O(e^{-Ck})$ for $\theta = 1/2$ were proved. These results, which can also be proven for the classic AFW with the same techniques, interpolate between the already well known $O(1/k)$ rate in the general convex case and the linear rate of the strongly convex one. Analogously, in [44] Holderian error bounds were used to interpolate between the classic FW method $O(1/k)$ rate for general convex functions and the $O(1/k^2)$ rate for strongly convex functions on strongly convex sets.
We now recall a few relevant facts about error bounds. For a more exhaustive reference with many examples of applications we refer the reader to [8].
A very important class of functions satisfying error bounds is that of semialgebraic functions, for which the Łojasiewicz's inequality implies an Holderian error bound condition.

Theorem 2.3.1 (Łojasiewicz's inequality)**.** *Let Ω be a closed and bounded semialgebraic set, f and g two continuous semialgebraic functions from Ω to $\mathbb{R}$ such that $f^{-1}(0) \subset g^{-1}(0)$. Then there exists an integer $N > 0$ and a constant $c \in \mathbb{R}$ such that $|g|^N \leq c|f|$ on Ω.*

This is a classical result and a reference is for instance [7], Corollary 2.6.7. We have the following corollary:

Corollary 2.3.2. *Let Ω be a compact semialgebraic set and $f : \Omega \to \mathbb{R}$ a semialgebraic continuous function. Let $S = \text{argmin} f$ and $f^* = \min f$. Then for some $N \in \mathbb{N}$ and $\gamma_0 > 0$*

$$\gamma_0 \text{dist}(x, S)^N \leq f(x) - f^* \tag{2.3.5}$$

for every $x \in \Omega$.

Proof. Since S is a semialgebraic set $\text{dist}(x, S) : \Omega \to \mathbb{R}$ is a semialgebraic function. Now let $\bar{f} = f(x) - f^*$. We have $\text{dist}(x, S)^{-1}(0) = S = \bar{f}^{-1}(0)$. We can then apply Łojasiewicz's inequality to $\text{dist}(x, S)$ and $\bar{f}(x)$ and obtain (2.3.5) with γ_0 for instance equal to $\frac{1}{c+1}$. □

When f is a convex piecewise polynomial we have the following Holderian error bound condition on sublevel sets:

Theorem 2.3.3. *([8], Proposition 8) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a piecewise convex polynomial with degree d. Suppose that $\text{argmin} f \neq \emptyset$. Then, for any $r \geq \inf f$ there exists $\gamma_r > 0$ such that*

$$\gamma_r \text{dist}(x, \text{argmin} f)^N \leq f(x) - f^*$$

for every $x \in \Omega_r = \{x \mid f(x) \leq r\}$ and for $N = (\deg(f) - 1)^{n+1} + 1$.

In order to study the connection between error bounds and $\nabla f|_\Omega$ we need to introduce the restriction $f_\Omega : \mathbb{R}^n \to (-\infty, \infty]$ of f to Ω given by

$$f_\Omega(x) = f(x) + i_\Omega(x) \tag{2.3.6}$$

where i_Ω is the indicator function of Ω:

$$i_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{if } x \notin \Omega \end{cases} \tag{2.3.7}$$

The function $f_\Omega(x)$ is of course not differentiable in $\mathbb{R}^n$ even when f is, but we can still relate its subgradient to error bounds using the subgradient norm:

Definition 2.3.4. For every $x \in \Omega$ the norm of the subgradient $\partial f_\Omega(x)$ is defined as

$$\|\partial f_\Omega(x)\| = \min_{y \in \partial f_\Omega(x)} \|y\| \tag{2.3.8}$$

The following proposition relates $\|\partial f_\Omega(x)\|$ to $-\nabla f(x)$ using the tangent cone $T_\Omega(x)$ to Ω.

Proposition 2.3.5. *For every $x \in \partial\Omega$ we have*

$$\|\partial f_\Omega(x)\| = \|\pi(T_\Omega(x), -\nabla f(x))\| \tag{2.3.9}$$

Proof. Since $\partial i_\Omega(x)$ is equal to the normal cone $N_\Omega(x)$ and $f_\Omega(x) = i(x) + f(x)$ we have, by the calculus rules for the subdifferential, $\partial f_\Omega(x) = \partial i(x) + \nabla f(x)$ so that

$$\partial f_\Omega(x) = N_\Omega(x) + \nabla f(x) = N_\Omega(x) - (-\nabla f(x))$$

Thus $\|\partial f_\Omega(x)\| = \min_{y \in N_\Omega(x)} \|y - (-\nabla f(x))\| = \mathrm{dist}(N_\Omega(x), -\nabla f(x))$. Since $T_\Omega(x) = N_\Omega(x)^d$ the conclusion follows from Lemma 5.1.10. $\square$

We can now prove an important relation between Holderian error bounds and a condition on $\nabla f(x)$.

Proposition 2.3.6. *Let* $0 < \theta \leq 1$.
For every $x \in \Omega$ *if* $\mathrm{dist}(x, \mathrm{argmin} f|_\Omega) \leq M(f(x) - f^*)^\theta$ *then*

$$\|\pi(T_\Omega(x), -\nabla f(x))\| \geq \frac{(f(x) - f^*)^{1-\theta}}{M}$$

Proof. Let x^* be the projection of x in $\mathrm{argmin} f|_\Omega$ so that $\|x - x^*\| = \mathrm{dist}(x, \mathrm{argmin} f|_\Omega)$ and $f(x^*) = f^*$. Then on the one hand by Lemma 5.1.10 we have

$$\frac{(-\nabla f(x), x^* - x)}{\|x^* - x\|} \leq \|\pi(T_\Omega(x), -\nabla f(x))\| \tag{2.3.10}$$

where by the definition of x^*

$$\frac{(-\nabla f(x), x^* - x)}{\|x^* - x\|} = \frac{(-\nabla f(x), x^* - x)}{\mathrm{dist}(x, \mathrm{argmin} f|_\Omega)} \tag{2.3.11}$$

On the other hand by convexity

$$(-\nabla f(x), x^* - x) \geq f(x) - f(x^*) = f(x) - f^* \tag{2.3.12}$$

so that

$$\frac{(-\nabla f(x), x^* - x)}{\mathrm{dist}(x, \mathrm{argmin} f|_\Omega)} \geq \frac{f(x) - f^*}{\mathrm{dist}(x, \mathrm{argmin} f|_\Omega)} \tag{2.3.13}$$

Applying the error bound hypothesis we obtain

$$\frac{f(x) - f^*}{\mathrm{dist}(x, \mathrm{argmin} f|_\Omega)} \geq \frac{(f(x) - f^*)^{1-\theta}}{M} \tag{2.3.14}$$

The conclusion follows concatenating all the inequalities we proved from (2.3.14) to (2.3.11) with the exception of (2.3.12). $\square$

The proposition above will allow us to use error bounds to prove converge rates for the FDFW under some assumptions on Ω.

2.4 The normalized width $\mathrm{NW}(\Omega)$

2.4.1 Motivation

For both step size strategies described in section 2.2 if the step k of a FDFW method is non maximal it is easy to prove using the standard descent lemma

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}(-\nabla f(x_k), \frac{d_k}{\|d_k\|})^2 \tag{2.4.1}$$

In the unconstrained case the choice that maximizes the right hand side is the gradient descent one, $d_k = -\nabla f(x_k)$. Instead if we are constrained to pick feasible directions we have

$$\sup_{y \in \Omega} \frac{(-\nabla f(x_k), y - x_k)}{\|y - x_k\|} = \|\pi(T_\Omega(x_k), -\nabla f(x_k))\| \tag{2.4.2}$$

(see Lemma 5.1.10 for a proof), and therefore

$$(-\nabla f(x_k), \frac{d_k}{\|d_k\|}) \leq \|\pi(T_\Omega(x_k), -\nabla f(x_k))\| \tag{2.4.3}$$

For the directions selected by the FDFW we want to prove that

$$(-\nabla f(x_k), \frac{d_k}{\|d_k\|}) \geq \bar{h}\|\pi(T_\Omega(x_k), -\nabla f(x))\| \tag{2.4.4}$$

for some fixed $0 < \bar{h} \leq 1$ and for every $x_k \in \Omega$. Since by Lemma 2.3.6

$$\|\pi(T_\Omega(x_k), -\nabla f(x))\| \geq \frac{(f(x) - f^*)^{1-\theta}}{M}$$

under the Holderian error bound condition (2.3.3) we can then use (2.4.4) to give a lower bound for the decrease of the objective function at every step.
This motivates the definition of normalized width.

2.4.2 The normalized width.

We will first define a parameter $\mathrm{NW}^f(\Omega)$ depending on f, Ω such that inequality (2.4.4) holds with $\bar{h} = \mathrm{NW}^f(\Omega)$ for the FDFW with away directions. We will then eliminate the dependence from f and define the normalized width $\mathrm{NW}(\Omega)$ depending only on the geometry of the set. This normalized width $\mathrm{NW}(\Omega)$ will be a lower bound for $\mathrm{NW}^f(\Omega)$, or in other words $\mathrm{NW}^f(\Omega) \geq \mathrm{NW}(\Omega)$ for every convex f differentiable in Ω.
Let

$$m_r = \mathrm{argmax}\{(r, v) \mid v \in \Omega\}$$

and

$$M_r(\bar{x}) = \mathrm{argmin}\{(r, v) \mid v \in \mathcal{F}(\bar{x})\}$$

Let

$$\sigma_r^*(\bar{x}) = \inf\{\frac{(r, s-q)}{\|s - x\| + \|q - x\|} \mid s \in m_r,\ q \in M_r(\bar{x})\} \tag{2.4.5}$$

Let $\pi_{\bar{x}}(r)$ be the norm of the projection of r on the tangent cone to Ω in $\bar{x}$:

$$\pi_{\bar{x}}(r) = \max(0, \sup_{h \in \Omega / \{\bar{x}\}} (r, \frac{h - \bar{x}}{\|h - \bar{x}\|})) = \|\pi(T_\Omega(\bar{x}), r)\|$$

If $\pi_{\bar{x}}(r) \neq 0$ we define the directional normalized width of Ω in $\bar{x}$ as

$$\mathrm{dirNW}(\Omega, \bar{x}, r) = \frac{\sigma_r^*(\bar{x})}{\pi_{\bar{x}}(r)}$$

Notice that when $\pi_{\bar{x}}(r) \neq 0$ we have $\bar{x} \notin m_r$ by first order optimality conditions so that the term $\|s - x\|$ is bounded away from zero in (2.4.5). Therefore by continuity and compactness the inf is actually a min and there exists $s_r^* \in m_r$, $q_r^* \in M_r(\bar{x})$ such that

$$\frac{(r, s_r^* - q_r^*)}{\|s_r^* - x\| + \|q_r^* - x\|} = \inf\{\frac{(r, s - q)}{\|s - x\| + \|q - x\|} \mid s \in m_r,\ q \in M_r(\bar{x})\} \tag{2.4.6}$$

Finally we define $\mathrm{NW}^f(\Omega)$ as

$$\mathrm{NW}^f(\Omega) = \inf_{\bar{x} \in \Omega \setminus X} \mathrm{dirNW}(\Omega, \bar{x}, -\nabla f(\bar{x})) \tag{2.4.7}$$

and

$$\mathrm{NW}(\Omega) = \inf_{\substack{\bar{x} \in \Omega, \\ r : \pi_{\bar{x}}(r) \neq 0}} \mathrm{dirNW}(\Omega, \bar{x}, r) \tag{2.4.8}$$

As anticipated $\mathrm{NW}(\Omega)$ is a lower bound for $\mathrm{NW}^f(\Omega)$:

Proposition 2.4.1. *For any convex $f \in C^1(\Omega)$*

$$\mathrm{NW}^f(\Omega) \geq \mathrm{NW}(\Omega) \tag{2.4.9}$$

Proof. By first order optimality conditions

$$\pi(-\nabla f(\bar{x}), T_\Omega(x)) = 0 \iff \bar{x} \in X \tag{2.4.10}$$

so that

$$\begin{aligned} \mathrm{NW}^f(\Omega) = \inf_{\bar{x} \in \Omega \setminus X} \mathrm{dirNW}(\Omega, \bar{x}, -\nabla f(\bar{x})) &= \inf_{\substack{\bar{x} \in \Omega, \\ \pi(-\nabla f(\bar{x}), T_\Omega(\bar{x})) \neq 0}} \mathrm{dirNW}(\Omega, \bar{x}, -\nabla f(\bar{x})) \\ \leq \inf_{\bar{x} \in \Omega} \inf_{\substack{r \in \mathbb{R}^n, \\ \pi(r, T_\Omega(\bar{x})) \neq 0}} \mathrm{dirNW}(\Omega, \bar{\bar{x}}, r) = \mathrm{NW}(\Omega) \end{aligned}$$

□

As we will see later in this section $\mathrm{NW}(\Omega)$ is greater than 0 not only for polytopes but also for strictly convex sets satisfying a certain condition for every vertex. This condition fundamentally imposes a bound on the variation of the width of Ω measured from a vertex along directions at a fixed distance from the boundary of the tangent cone. This boundedness property holds for instance for all sets whose boundary looks locally like a sphere.

This is interesting because when $\mathrm{NW}(\Omega) > 0$ using the same technique employed in [33] for the AFW it is possible to prove linear convergence for the FDFW with away directions for strongly convex functions with Lipschitz gradient.
Unfortunately there exists strongly convex sets with $\mathrm{NW}(\Omega) = 0$ even in $\mathbb{R}^2$. Building a strongly convex set for which there actually is no linear convergence for the FDFW seems more complicated.

Remark 2.4.2. the main properties of $\mathrm{NW}(\Omega)$ still hold if we only consider the vertexes maximizing or minimizing the linear function r.

2.4.3 Bounds for $\mathrm{NW}(\Omega)$ on polytopes

We now prove a lower bound for $\mathrm{NW}(\Omega)$ using the facial distance $\Phi(\Omega)$ studied in [39], which is also equal to the pyramidal width $\mathrm{PWidth}(\Omega)$ studied in [34] (see Theorem 2, [39]). Before proving the lower bound on $\mathrm{NW}(\Omega)$ we briefly recall the definition of the facial distance $\Phi(\Omega)$.
Let $A = [a_1, ..., a_m] \in \mathbb{R}^{n\times m}$ a matrix with m columns; with a slight abuse of notation we use A also to denote the set $\{a_1, ..., a_m\} \subset \mathbb{R}^n$. For $x \in \Delta_{n-1}$, we define $I(x)$ as the set of indexes greater than 0:

$$I(x) = \{i \in \{1, ..., n\} : x_i > 0\} \tag{2.4.11}$$

and

$$S(x) = \{a_i : i \in I(x)\} \tag{2.4.12}$$

For $x, z \in \Delta_{n-1}$ such that $A(x - z) \neq 0$ let $d = A(x - z)/\|A(x - z)\|$. We finally define

$$\Phi(A, x, z) = \min_{p\in\mathbb{R}^m:(p,d)=1} \ \max_{s\in S(x),a\in A} (p, s - a), \tag{2.4.13}$$

and

$$\Phi(A) = \min_{x,z\in\Delta_{m-1}:A(x-z)\neq 0} \Phi(A, x, z) \tag{2.4.14}$$

The analogy between $\mathrm{dirNW}(\Omega, x, r)$ and $\Phi(A, x, z)$ is evident: in both definitions x is used as a "center" from which we compute a certain minimum considering all directions r with $\pi(T_\Omega(x), r) \neq 0$ and p with $(p, d) = 1$ for some d respectively.
We now prove an inequality relating these two quantities whenever $\Omega = \mathrm{conv}(A)$.

Proposition 2.4.3. *Let* $\Omega = \mathrm{conv}(A)$, *where* $A = \{a_1, ..., a_m\}$ *is a finite set of vectors in* $\mathbb{R}^n$, *and assume* $\mathrm{diam}(\Omega) = D$.
Then $\mathrm{NW}(\Omega) \geq \frac{\Phi(A)}{2D}$.

We begin with the following lemma:

Lemma 2.4.4. *Under the hypotheses of proposition* (2.4.3) *for every* $x \in \Delta_{m-1}$, $p \in \mathbb{R}^n \setminus \{0\}$ *there exists* $z \in \Delta_{m-1}$ *such that*

$$(p, A(x - z)) > 0 \tag{2.4.15}$$

if and only if

$$\pi(T_\Omega(Ax), -p) \neq 0 \tag{2.4.16}$$

Proof. By Lemma 5.1.10 for every $p \in \mathbb{R}^n \setminus \{0\}$ we have

$$\pi(T_\Omega(Ax), -p) = \max(0, \sup_{h \in \Omega \setminus \{Ax\}} \frac{(-p, h - Ax)}{\|h - Ax\|}) = \max(0, \sup_{\substack{z \in \Delta_{m-1}: \\ A(x-z) \neq 0}} \frac{(-p, A(z - x))}{\|A(z - x)\|}) \tag{2.4.17}$$

where we used $\Omega = \mathrm{conv}(A)$ to apply the substitution $h = Az$ in the last equality. The right hand side of (2.4.17) is greater than 0 if and only if (2.4.15) holds for some $z \in \Delta_{m-1}$, from which the desired equivalence follows. □

We now prove the result reported in Proposition 2.4.3.

Proof. In the rest of this proof

$$d = \frac{A(x - z)}{\|A(x - z)\|} \tag{2.4.18}$$

with $x, z \in \Delta_{m-1}$ and $A(x - z) \neq 0$. We have

$$\Phi(A, x, z) \stackrel{\text{def}}{=} \min_{p \in \mathbb{R}^m : (p,d)=1} \max_{q \in S(x), s \in A} (p, q - s) = \min_{p \in \mathbb{R}^m : (p,d)>0} \max_{q \in S(x), s \in A} \frac{(p, q - s)}{(p, d)}$$

Now $\Omega = \mathrm{conv}(A)$ by hypothesis and $S(x) \subseteq \mathcal{F}(Ax)$ because Ax is a proper combination of the elements in $S(x)$, which therefore are all in the minimal face of $\mathrm{conv}(A)$ containing Ax. We then have

$$\max_{q \in S(x), s \in A} \frac{(p, q - s)}{(p, d)} \leq \max_{q \in \mathcal{F}(Ax), s \in \Omega} \frac{(p, q - s)}{(p, d)} = \frac{(p, q_{-p} - s_{-p})}{(p, d)} \tag{2.4.19}$$

for any $q_{-p} \in \mathrm{argmax}\{(p, q) \mid q \in \mathcal{F}(Ax)\}$, $s_{-p} \in \mathrm{argmin}\{(p, s) \mid s \in \Omega\}$. Here we are considering $-p$ instead of p to keep the notation consistent with the one used in the definition of $\mathrm{NW}(\Omega)$ as it will be apparent later in the proof. Let Π_{Ax} be the set defined by

$$\Pi_{Ax} = \{r \in \mathbb{R}^n \setminus \{0\} \mid \exists z \in \Delta_{m-1} \text{ such that } (r, Az - Ax) > 0\}$$

By Lemma 2.4.15 we have

$$\Pi_{Ax} = \{p \in \mathbb{R}^n \setminus \{0\} \mid \pi(T_\Omega(Ax), -p) \neq 0\} \tag{2.4.20}$$

We now rewrite $\Phi(A)$ switching two minimization operators.

$$\begin{aligned} \Phi(A) &= \min_{\substack{x,z \in \Delta_{m-1}: \\ A(x-z) \neq 0}} \Phi(A, x, z) = \min_{\substack{x,z \in \Delta_{m-1}: \\ A(x-z) \neq 0}} \min_{\substack{p \in \mathbb{R}^m: \\ (p,d)>0}} \frac{(p, s_{-p} - q_{-p})}{(p, d)} \\ &= \min_{\substack{x \in \Delta_{m-1}, \\ p \in \Pi_{Ax}}} \min_{\substack{z \in \Delta_{m-1}: \\ (p,d)>0}} \frac{(p, s_{-p} - q_{-p})}{(p, d)} = \min_{\substack{x \in \Delta_{m-1} \\ p \in \Pi_{Ax}}} \min_{\substack{z \in \Delta_{m-1}: \\ (p,d)>0}} \frac{(p, s_{-p} - q_{-p})}{(-p, -d)} \end{aligned} \tag{2.4.21}$$

We are now going to eliminate Δ_{m-1} from this expression. First we observe that since $\Omega = \mathrm{conv}(A)$

$$\{d \in \mathbb{R}^n \mid d = \frac{A(x - z)}{\|A(x - z)\|}, z \in \Delta_{m-1}\} = \{\frac{h - Ax}{\|h - Ax\|} \mid h \in \Omega,\ h \neq Ax\}$$

Then

$$\begin{aligned}
&\min_{\substack{x\in\Delta_{m-1}\\ p\in\Pi_{Ax}}}\ \min_{\substack{z\in\Delta_{m-1}:\\ (p,d)>0}} \frac{(p, s_{-p}-q_{-p})}{(-p,-d)} = \min_{\substack{\bar{x}\in\Omega\\ p\in\Pi_{\bar{x}}}}\ \min_{\substack{h\in\Omega,\\ (-p,h-\bar{x})>0}} \frac{(p, s_{-p}-q_{-p})\|h-\bar{x}\|}{(-p,h-\bar{x})} \\
= &\min_{\substack{\bar{x}\in\Omega\\ r:\pi(T_\Omega(\bar{x}),r)\neq 0}}\ \min_{h\in\Omega,\ (r,h-\bar{x})>0} \frac{(r, q_r - s_r)\|h-\bar{x}\|}{(r,h-\bar{x})}
\end{aligned} \tag{2.4.22}$$

where we used $\Omega = \text{conv}(A)$ to apply the substitution $d = \frac{Ax-Az}{\|Ax-Az\|} = \frac{\bar{x}-h}{\|x-h\|}$ in the first equality. Now as a direct consequence of Lemma 5.1.10

$$\min_{h\in\Omega\setminus\{Ax\}} \frac{\|h-Ax\|}{(r,h-Ax)} = 1/\max_{h\in\Omega\setminus\{Ax\}} \frac{(r,h-Ax)}{\|h-Ax\|} = \frac{1}{\|\pi(T_\Omega(Ax),r)\|} \tag{2.4.23}$$

for every r such that $\|\pi(T_\Omega(Ax),r)\| \neq 0$.
Therefore

$$\begin{aligned}
\min_{\substack{\bar{x}\in\Omega\\ r:\pi(T_\Omega(x),r)\neq 0}}\ \min_{\substack{h\in\Omega,\\ (r,h-Ax)>0}} \frac{(r, q_r - s_r)\|h-\bar{x}\|}{(r,h-\bar{x})} &= \min_{\bar{x}\in\Omega}\ \min_{r:\pi(T_\Omega(\bar{x}),r)\neq 0} \frac{(r,q_r-s_r)}{\|\pi(T_\Omega(\bar{x}),r)\|} \leq \\
&\leq \min_{\bar{x}\in\Omega} 2D\text{dirNW}(\Omega,\bar{x},r)
\end{aligned} \tag{2.4.24}$$

where the last inequality follows immediately from the definition of $\text{dirNW}(\Omega,\bar{x},r)$ and $D = \text{diam}(\Omega)$:

$$2D\text{dirNW}(\Omega,\bar{x},r) \geq \text{dirNW}(\Omega,\bar{x},r)(\|q_r^*-\bar{x}\|+\|s_r^*-\bar{x}\|) = \frac{(r,q_r^*-s_r^*)}{\|\pi(T_\Omega(\bar{x}),r)\|} = \frac{(r,q_r-s_r)}{\|\pi(T_\Omega(\bar{x}),r)\|} \tag{2.4.25}$$

Concatenating (2.4.21), (2.4.22), (2.4.24) we get

$$\Phi(A) \leq \min_{\bar{x}\in\Omega} 2D\text{dirNW}(\Omega,\bar{x},r) = 2D\text{NW}(\Omega) \tag{2.4.26}$$

□

We now give a lower bound for $\text{NW}(\Omega)$ on polytopes as a function of simple geometric properties of the polytope, using a result proved in [39] for $\Phi(A)$.

Corollary 2.4.5. *Assume* $\Omega = \text{conv}(A)$ *with* $A = \{a_1,...,a_m\} \subset \mathbb{R}^n$, $m \geq 2$ *and let* $D = \text{diam}(\Omega)$. *Then*

$$\text{NW}(\Omega) \geq \min_{\substack{F\in\text{faces}(\text{conv}(A))\\ \emptyset\subsetneq F\subsetneq\text{conv}(A)}} \text{dist}(F, \text{conv}(A\setminus F))/2D \tag{2.4.27}$$

Furthermore, if $F \in \text{faces}(\text{conv}(A))$ *minimizes the right hand side, then there exists* $h \in \text{conv}(S)$, *with* $S \subset A \setminus F$, $\bar{x} \in F$ *such that*

$$h \in \text{argmax}_{y\in\text{conv}(S)}(h-\bar{x},y),\ \ \bar{x} \in \text{argmin}_{y\in\Omega}(h-\bar{x},y) \tag{2.4.28}$$

and

$$\text{NW}(\Omega) \geq \frac{\|h-\bar{x}\|}{2D} \tag{2.4.29}$$

Proof. By [39], Theorem 1 the right hand side in (2.4.27) is equal to $\Phi(A)$, so that (2.4.27) is equivalent to Proposition 2.4.3 we just proved.
Again by [39], Theorem 1 there exists $x, z \in \Delta_{m-1}$ such that $Az \in F$, $Ax \in \text{conv}(A \setminus F)$ and

$$\Phi(A) = \max_{\substack{s \in S(x) \\ a \in \text{conv}(A)}} (p, s - a) = \|Ax - Az\| \tag{2.4.30}$$

where $p = \frac{A(x-z)}{\|A(x-z)\|}$. Let $h = Az$ and $\bar{x} = Ax$, so that

$$p = \frac{\bar{x} - h}{\|\bar{x} - h\|} \tag{2.4.31}$$

We can then write

$$\max_{y \in \text{conv}(S)} (p, y) - \min_{a \in \Omega} (p, a) = \max_{\substack{y \in \text{conv}(S) \\ a \in \Omega}} (p, y - a) = \Phi(A) = (p, Ax - Az) = (p, \bar{x} - h) = (p, \bar{x}) - (p, h) \tag{2.4.32}$$

Since $\bar{x} \in \text{conv}(S)$ and $h \in \Omega$ equating the first and the last term of (2.4.32) it follows necessarily (2.4.28). Finally

$$\frac{\|h - \bar{x}\|}{2D} = \frac{\|Ax - Az\|}{2D} = \frac{\Phi(A)}{2D} \leq \text{NW}(\Omega) \tag{2.4.33}$$

$\square$

2.4.4 Bounds for NW(Ω) on strictly convex sets

In this section we define sufficient conditions for NW(Ω) to be greater than 0 on strictly convex sets. In particular we will prove that NW(Ω) is greater than 0 on sets whose boundary looks locally like a sphere. Remarkably, this particular hypothesis on the boundary has already been used in a weaker form in [20] to prove linear convergence for the classic FW method with the additional assumption of a unique non singular minimum. This result together with the work in [19] about perturbations on the feasible set provided a theoretical justification for the effectiveness of the FW algorithm in highly constrained problems.
When computing NW(Ω) one can always assume that Ω is full dimensional. This is not restrictive since NW(Ω) does not depend on the dimension of the space containing Ω, as a corollary of the following proposition:

Proposition 2.4.6. *Assume that* $T : \text{aff}(\Omega) \to \mathbb{R}^{\dim(\text{aff}(\Omega))}$ *is an isometry. Then* $\text{NW}(\Omega) = \text{NW}(T(\Omega))$.

Proof. For a fixed $\bar{x}$ in Ω and for every $r \in \mathbb{R}^n$ we have $(r, z - \bar{x}) = (\pi(\text{aff}(\Omega), r), z - \bar{x})$ for every $z \in \Omega$, so that $\text{dirNW}(\Omega, \bar{x}, r) = \text{dirNW}(\Omega, \bar{x}, \pi(\text{aff}(\Omega), r))$.Therefore in the definition of NW the inf is taken on the same sets for Ω and $T(\Omega)$. $\square$

We now introduce the directional length function

Definition 2.4.7. For $\bar{x} \in \Omega$ we define $l_{\Omega,\bar{x}} : T_\Omega(\bar{x}) \to \mathbb{R}_{\geq 0}$ as

$$l_{\Omega,\bar{x}}(c) = \sup\{k \in \mathbb{R} \mid \bar{x} + kc \in \Omega\} \tag{2.4.34}$$

This function is the fundamental block to define bounds for the $\mathrm{NW}(\Omega)$ on strictly convex sets. A few of its properties are proved in the appendix. Most notably, in the interior of $T_\Omega(\bar{x})$ it is the pointwise inverse of the Minkowski functional of $\Omega - \{\bar{x}\}$. We remark that since we are assuming Ω compact the sup in (2.4.34) is actually a max for $c \neq 0$.
The most general result in this section is that we can give a lower bound on $\mathrm{NW}(\Omega)$ using another geometric parameter which is possible to compute knowing the local behaviour of Ω. Before stating the result it is convenient to define the main building blocks of this parameter.

Definition 2.4.8. Let Ω be a compact strictly convex set and $\bar{x} \in \partial\Omega$. We define the following for $0 \leq \beta \leq \delta \leq 1$.

$$\begin{aligned} l^B_{\Omega,\bar{x}}(\beta,\delta) &= \sup\{l_{\Omega,\bar{x}}(\hat{c}) \mid \beta \leq \mathrm{dist}(\hat{c}, T_\Omega(\bar{x})^c) \leq \delta\} \\ l^b_{\Omega,\bar{x}}(\beta,\delta) &= \inf\{l_{\Omega,\bar{x}}(\hat{c}) \mid \beta \leq \mathrm{dist}(\hat{c}, T_\Omega(\bar{x})^c) \leq \delta\} \end{aligned} \tag{2.4.35}$$

Finally, for $0 < k \leq 1$ we define:

$$R_{\Omega,\bar{x}}(k) = \inf_{0<\delta\leq 1} \frac{l^b_{\Omega,\bar{x}}(k\delta,\delta)}{l^B_{\Omega,\bar{x}}(0,\delta)} \tag{2.4.36}$$

Some motivation and properties of these functions are reported in the appendix. We can now state the main theorem:

Theorem 2.4.9. *Let Ω be a strictly convex set and let $k = \alpha/D$, with $\alpha > 0$ such that there exists a ball of radius α contained in Ω. Assume that:*

$$\inf_{\bar{x}\in\Omega} R_{\Omega,\bar{x}}(k) = M > 0 \tag{2.4.37}$$

for some $k \leq \alpha/4D$.
Then if Ω has width W and diameter D

$$\mathrm{NW}(\Omega) \geq \min\{\frac{W}{2D}, \frac{M}{2}\}$$

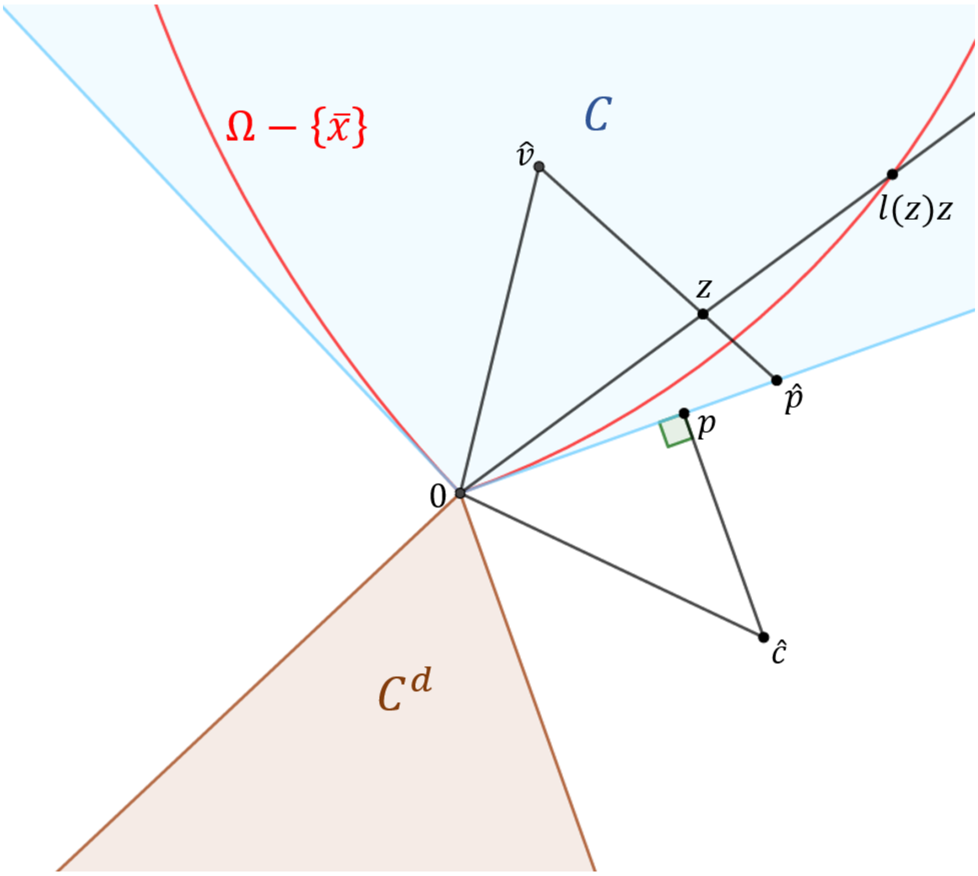
We now introduce some notation to present an outline of the proof. We fix $\bar{x} \in \Omega$ so that for simplicity we can write C instead of $T_\Omega(\bar{x})$ and l instead of $l_{\Omega,\bar{x}}$. We will always use c to represent a generic vector in $\mathbb{R}^n/C^d$, with $\{s(c)\} = \mathrm{argmax}\{(c,y) \mid y \in \Omega\}$ and $\{q(c)\} = \mathrm{argmin}\{(c,y) \mid y \in \Omega\}$, where the argmax and the argmin are singletons for the strict convexity of Ω. Since $\mathrm{dirNW}(\Omega,\bar{x},c)$ does not change if we multiply c by a positive scalar we can always consider $\hat{c}$ instead of c. The point $p \in C \setminus \{0\}$ will be the projection of $\hat{c}$ on C.
Even if the proof is rather long and technical, the main ideas are quite simple. In order to bound $\mathrm{NW}(\Omega)$ we need to bound $\mathrm{dirNW}(\Omega,\bar{x},c)$ for every $\bar{x} \in \Omega$ and c such that $\pi(T_\Omega(\bar{x}),c) \neq 0$. We will distinguish two cases according to whether $\bar{x} \in \Omega^\circ$ or $\bar{x} \in \partial\Omega$. In the first case it is rather straightforward to prove $\mathrm{dirNW}(\Omega,\bar{x},c) \geq W/2D$ without using the hypothesis on $R_{\Omega,\bar{x}}$. When $\bar{x} \in \partial\Omega$ the proof is more technical. In this case the expression for $\mathrm{dirNW}(\Omega,\bar{x},c)$ simplifies as

$$\mathrm{dirNW}(\Omega,\bar{x},c) = \frac{(c, s(c) - \bar{x})}{\|p\|\|\bar{x} - s(c)\|} \tag{2.4.38}$$

In order to give a lower bound to this quantity we first give an upper bound on $\|\bar{x}-s(c)\|$ in terms of l^B using a property of cones; then we give a lower bound on $\frac{(c,s(c)-\bar{x})}{\|p\|}$ in terms of l^b identifying a point $\tilde{s}$ in the form of $\bar{x}+zl(z)$ such that $(c,(\bar{x}+zl(z))-\bar{x})=(c,zl(z))$ is "large enough". The theorem will finally follow applying hypothesis (2.4.37).

In the rest of the proof $\delta^*(\bar{x})=\max\{\mathrm{dist}(\hat{d},T_\Omega(\bar{x})^c)/\{0\}\mid d\in T_\Omega(\bar{x})\setminus\{0\}\}$ is the

Figure 2.1: Configuration of the proof when $\bar{x}$ is on $\partial\Omega$.



maximum distance of a unitary vector in $T_\Omega(\bar{x})$ from the border of the cone. We will use δ^* instead of $\delta^*(\bar{x})$ when $\bar{x}$ will be obvious from the context.

Proof. Since Ω contains a ball of radius α we have $W>2\alpha$ by monotony of the width function.

Furthermore, if $q+\bar{x}\in\Omega$ such that $B(q+\bar{x},\alpha)\subset\Omega$ then $B(q,\alpha)\subset\Omega-\{x\}\subset C$ and as a consequence

$$\delta^*\geq\mathrm{dist}(\hat{q},C^c)=\frac{1}{\|q\|}\mathrm{dist}(q,C^c)\geq\frac{1}{D}\mathrm{dist}(q,C^c)\geq\frac{\mathrm{dist}(q,(\Omega-\{x\})^c)}{D}\geq\frac{\alpha}{D} \tag{2.4.39}$$

We now distinguish two cases according to the position of $\bar{x}$.

Case 1: $\bar{x}\in\Omega^\circ$. Under this hypothesis $(c,s(c)-q(c))$ is simply the directional width of Ω with respect to c, which is at most $\|c\|W$. We also have $\pi_{\bar{x}}(c)=\|c\|$ since

$\bar{x} \in \Omega^\circ$. Putting together these properties we can conclude

$$\mathrm{dirNW}(\Omega, \bar{x}, c) = \frac{(c, s(c) - q(c))}{\pi_{\bar{x}}(c)(\|s(c) - \bar{x}\| + \|q(c) - \bar{x}\|)} \geq \frac{W\|c\|}{\|c\|(\|s(c) - \bar{x}\| + \|q(c) - \bar{x}\|)} \geq \frac{W}{2D} \tag{2.4.40}$$

Case 2: $\bar{x} \in \partial\Omega$. Let $\delta = \mathrm{dist}(\hat{c}, C^d)$ and s be the point maximizing c in Ω. Then $(c, s - \bar{x}) > 0$ so that by Proposition 5.1.4 $\hat{c} \in C^d_\delta$ implies that $s - \bar{x}$ is not in $C_{-\delta}$ which means $\mathrm{dist}(\hat{c}, C^c) < \delta$. Therefore

$$\|s - \bar{x}\| \leq \max\{l(\hat{w}) \mid \mathrm{dist}(\hat{w}, C^c) < \delta\} \tag{2.4.41}$$

We now recall that p is the projection of $\hat{c}$ in C, so that $\|p\| = \delta$ by Lemma 5.1.2. By Proposition 5.2.3 there exists $\hat{v} \in C \cap -C^d$ such that $\mathrm{dist}(\hat{v}, C^c) = \delta^*$. Consider the point

$$z = (1 - \frac{\delta}{4})\hat{p} + \frac{\delta}{4}\hat{v}$$

We want to show that $(c, \cdot)$ decreases enough along the ray $\bar{x} + \lambda z$, $\lambda \geq 0$ before the ray reaches the boundary of Ω.
By the concavity of C^c

$$\mathrm{dist}(z, C^c) = \mathrm{dist}((1 - \frac{\delta}{4})\hat{p} + \frac{\delta}{4}\hat{v}, C^c) \geq (1 - \frac{\delta}{4})\mathrm{dist}(\hat{p}, C^c) + \frac{\delta}{4}\mathrm{dist}(\hat{v}, C^c) = \delta^*\delta/4 \tag{2.4.42}$$

and by convexity of the norm

$$\|z\| \leq (1 - \frac{\delta}{4})\|\hat{p}\| + \frac{\delta}{4}\|\hat{v}\| \leq 1 \tag{2.4.43}$$

We can now use (2.4.42) and (2.4.43) to give a lower bound for $\mathrm{dist}(\hat{z}, C^c)$

$$\mathrm{dist}(\hat{z}, C^c) = \mathrm{dist}(z, C^c)/\|z\| \geq \delta^*\delta/4 \tag{2.4.44}$$

Again by convexity

$$\mathrm{dist}(z, \mathrm{conv}(0, \hat{p})) = \mathrm{dist}((1 - \frac{\delta}{4})\hat{p} + \frac{\delta}{4}\hat{v}, \mathrm{conv}(\{0, \hat{p}\}) \leq \frac{\delta}{4}\mathrm{dist}(\hat{v}, \mathrm{conv}(\{0, \hat{p}\})) \leq \frac{\delta}{4} \tag{2.4.45}$$

By Proposition 5.2.3 we have $(p, v) \geq 0$ so that

$$\|z\| \geq \|(1 - \frac{\delta}{4})\hat{p} + \frac{\delta}{4}\hat{v}\| \geq ((1 - \frac{\delta}{4})^2 + (\frac{\delta}{4})^2)^{\frac{1}{2}} \geq \frac{1}{\sqrt{2}} > \frac{1}{2} \tag{2.4.46}$$

and therefore

$$\mathrm{dist}(\hat{z}, C^c) = \mathrm{dist}(z, C^c)/\|z\| < 2\mathrm{dist}(z, C^c) \leq 2\mathrm{dist}(z, \mathrm{conv}(0, \hat{p})) \leq \frac{\delta}{2} \tag{2.4.47}$$

where the second inequality is justified by $0, \hat{p} \in \partial C$ and we used (2.4.45) in the last inequality.

Now since p is the projection of $\hat{c}$ on the cone C we have $\hat{c} = p + p^{\perp}$, with $(p^{\perp}, p) = 0$. We apply this to bound $(\hat{c}, \hat{z})$:

$$\begin{aligned}(\hat{c},\hat{z}) =& \frac{1}{\|z\|}(\hat{c}, z) = \frac{1}{\|z\|}(\hat{c}, \hat{p} + \frac{\delta}{4}(\hat{v} - \hat{p})) = \frac{1}{\|z\|}((p + p^{\perp}, \hat{p}) + \frac{\delta}{4}(\hat{c}, \hat{v} - \hat{p})) \geq \\ \geq& \frac{1}{\|z\|}(\|p\| - \frac{\delta}{4}\|\hat{p} - \hat{v}\|) = \frac{1}{\|z\|}(\delta - \frac{\delta}{4}\|\hat{p} - \hat{v}\|) \geq \frac{1}{\|z\|}(\delta - 2\frac{\delta}{4}) \geq \frac{\delta}{2\|z\|} \geq \frac{\delta}{2}\end{aligned} \tag{2.4.48}$$

By equations (2.4.47) and (2.4.44) we have

$$\delta^* \frac{\delta}{4} \leq \operatorname{dist}(\hat{z}, C^c) \leq \frac{\delta}{2} \tag{2.4.49}$$

which implies

$$l(\hat{z}) \geq \min\{l(\hat{w}) \mid \delta\delta^*/4 \leq \operatorname{dist}(\hat{w}, C^c) \leq \delta/2\}$$

We give a lower bound for $(s - \bar{x}, \hat{c})$ considering the point $\bar{x} + zl(z) = \bar{x} + \hat{z}l(\hat{z}) \in \Omega$:

$$(s - \bar{x}, \hat{c}) = \max_{y \in \Omega}(y - \bar{x}, \hat{c}) \geq (l(\hat{z})\hat{z}, \hat{c}) \geq \frac{\delta}{2} l(\hat{z}) \tag{2.4.50}$$

where we applied (2.4.48) to bound $(\hat{z}, \hat{c})$ in the last inequality.
Notice that with these relations together with (2.4.41) we have lower and upper bounds for every term appearing in the computation of $\operatorname{dirNW}(\Omega, \bar{x}, c)$.
Indeed, we can write

$$\begin{aligned}\operatorname{dirNW}(\Omega, \bar{x}, c) =& \frac{(\hat{c}, (q - \bar{x}))}{(\|p\|\|q - \bar{x}\|)} \geq \frac{\delta/2 l(\hat{z})}{\delta\|q - \bar{x}\|} \geq \frac{1}{2}\frac{\min\{l(\hat{w}) \mid \delta^*\delta/4 \leq \operatorname{dist}(\hat{w}, C^c) \leq \delta/2\}}{\max\{l(\hat{w}) \mid \operatorname{dist}(\hat{w}, C^c) \leq \delta\}} \geq \\ \geq& \frac{1}{2}\frac{\min\{l(\hat{w}) \mid \delta^*\delta/4 \leq \operatorname{dist}(\hat{w}, C^c) \leq \delta\}}{\max\{l(\hat{w}) \mid \operatorname{dist}(\hat{w}, C^c) \leq \delta\}} \geq \frac{1}{2} R_{\Omega,\bar{x}}(\delta^*/4)\end{aligned} \tag{2.4.51}$$

and since $R_{\Omega,\bar{x}}$ is monotone increasing with $\delta^*/4 \geq \alpha/4D$ by (2.4.39)

$$\frac{1}{2} R_{\Omega,\bar{x}}(\delta^*/4) \geq \frac{1}{2} R_{\Omega,\bar{x}}(\alpha/4D) \tag{2.4.52}$$

We can now prove the thesis for $\bar{x} \in \partial\Omega$ taking the inf on both sides of inequality (2.4.51)

$$\begin{aligned}\inf_{\substack{\bar{x} \in \partial\Omega \\ c : \pi(T_\Omega(\bar{x}), c) \neq 0}} \operatorname{dirNW}(\Omega, \bar{x}, c) \geq& \inf_{\bar{x} \in \partial\Omega} \frac{1}{2} R_{\Omega,\bar{x}}(\delta^*(\bar{x})/4) \geq \\ \geq& \frac{1}{2} \inf_{\bar{x} \in \partial\Omega} R_{\Omega,\bar{x}}(\alpha/4D) \geq \frac{1}{2} \inf_{\bar{x} \in \partial\Omega} R_{\Omega,\bar{x}}(k) \geq \frac{M}{2}\end{aligned} \tag{2.4.53}$$

where we used hypothesis (2.4.37) in the last inequality. Combining (2.4.53) with inequality (2.4.40) we proved for case 1 we can finish the proof

$$\operatorname{NW}(\Omega) = \inf_{\substack{\bar{x} \in \Omega \\ c : \pi(T_\Omega(\bar{x}), c) \neq 0}} \operatorname{dirNW}(\Omega, \bar{x}, c) \geq \min(\frac{M}{2}, \frac{W}{2D}) \tag{2.4.54}$$

□

Remark 2.4.10. A possibly weaker hypothesis alternative to (2.4.37) can be $\inf R_{\Omega,\bar{x}}(\delta^*(\bar{x})/4) = M > 0$.

Remark 2.4.11. In the proof above the largest possible α is the inradius of Ω, which is the sup among the radii of the balls contained in Ω. As proved for instance in [26], there always exists a ball inscribed in Ω with radius equal to the inradius of Ω.

Remark 2.4.12. As anticipated in the introduction, there exists strongly convex subsets of $\mathbb{R}^2$ with $\mathrm{NW}(\Omega) = 0$. When Ω is not strictly convex, we conjecture that if there exists a face F of Ω of dimension $\leq \dim(\Omega) - 2$ such that around that face Ω is strictly convex, then $\mathrm{NW}(\Omega) = 0$. The intuition behind this claim comes from considering small directions orthogonal to F.

We now want to show that if Ω locally looks like a sphere, meaning that its second fundamental form is positive definite, we have $\mathrm{NW}(\Omega) > 0$. We will assume that Ω is smooth so that for every $x \in \Omega$ the tangent cone is a semi space, and there exists a unique tangent plane $T_x\Omega$. To analyze the local behaviour of Ω we will intersect it with compact cylinders $C^x_{r,\varepsilon}$ contained in $T_\Omega(x)$ with base $T_x\Omega \cap B(x,r)$ and height ε. More explicitly, if π_x is the projection on $T_x\Omega$

$$C^x_{r,\varepsilon} = \{y \in T_\Omega(x) \mid \|\pi_x(y) - x\| \leq r,\ \|y - \pi_x(y)\| \leq \varepsilon\} \tag{2.4.55}$$

Having introduced this key elements we can state the main result of this section.

Theorem 2.4.13. *Let $r, \varepsilon, h > 0$ be with the following property: for every $x \in \Omega$ the set $\partial\Omega \cap C^x_{r,\varepsilon}$ is the graph of a function $f : \bar{B}(x,\varepsilon) \to \mathbb{R}_{\geq 0}$ with respect to $T_x\Omega \cap \bar{B}(x,\varepsilon)$ with f such that $L_x h\|y\|^2 \leq f(y) \leq L_x\|y\|^2$ for every $y \in \bar{B}(0,\varepsilon)$ and for some $L_x > 0$. Then* $\mathrm{NW}(\Omega) > 0$.

To prove the theorem we first need to bound the width diameter ratio $R_{\Omega,\bar{x}}$ using the parameters we introduced. This is possible because thanks to the regularity conditions we imposed on Ω we can give a lower and an upper bound of Ω in the sense of inclusion. The lower and the upper bound on Ω will then translate in lower and upper bound for the length of Ω measured along a fixed direction from a vertex. Thanks to the radial symmetry of the bounds on Ω the upper and lower estimates for this kind of directional length depend only on the angle between the direction and the plane $T_x\Omega$. Therefore, we can use these estimates to bound $R_{\Omega,\bar{x}}$.
In the rest of this section we use $l_x(c)$, $l^b_x(\delta)$ and $l^B_x(\delta)$ as shorthands for $l_{\Omega,x}(c)$, $l^b_{\Omega,x}(\delta,\delta)$ and $l^B_{\Omega,x}(\delta,\delta)$ respectively.

Lemma 2.4.14. *Let D be the diameter of Ω, $\bar{x} \in \partial\Omega$. Under the hypotheses introduced in theorem* 2.4.13*:*

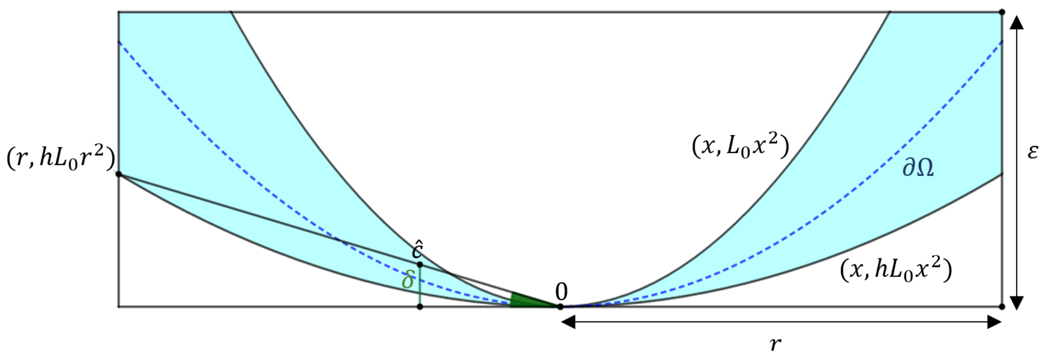
1. *For every* $0 < \delta \leq 1$*:*

$$l^b_{\bar{x}}(\delta) \geq \begin{cases} \frac{\delta}{L_{\bar{x}}(1-\delta^2)} & \text{for } \frac{\delta}{\sqrt{1-\delta^2}} \leq rL_{\bar{x}}h \\ \min(r, \varepsilon, \frac{\delta}{\sqrt{1-\delta^2}}) & \text{for } \frac{\delta}{L_{\bar{x}}(1-\delta^2)} > rL_{\bar{x}}h \end{cases} \tag{2.4.56}$$

2. *For δ such that*

$$l^B_{\bar{x}}(\delta) \leq \begin{cases} \frac{\delta}{L_{\bar{x}}h(1-\delta^2)} & \text{for } \frac{\delta}{\sqrt{1-\delta^2}} \leq rL_{\bar{x}}h \\ D & \text{for } \frac{\delta}{\sqrt{1-\delta^2}} > rL_{\bar{x}}h \end{cases} \tag{2.4.57}$$

Figure 2.2: $\partial\Omega$ behaves locally like the boundary of a sphere. The limit case of Lemma 2.4.14 with $\operatorname{dist}(\hat{c}, T_0\Omega) = s^{-1}(hL_0r)$ is represented.



3. *Let* $s : [0,1) \to \mathbb{R}_{\geq 0}$ *be defined by* $s(\delta) = \frac{\delta}{\sqrt{1-\delta^2}}$, $\delta_* = s^{-1}(rL_{\bar{x}}h)$. *For every* $0 < k \leq 1$

$$R_{\Omega,\bar{x}}(k) \geq \min\left(\frac{r}{D}, \frac{\varepsilon}{D}, \frac{k\delta_*}{L_{\bar{x}}(1-k^2\delta_*^2)}, \frac{kh(1-\delta_*)}{1-k\delta_*}\right) \tag{2.4.58}$$

Proof. Let $\hat{c}$ be in the semispace $T_\Omega(\bar{x})$ with $\delta = \operatorname{dist}(\hat{c}, T_{\bar{x}}\Omega) \in (0,1) \leq \delta$. For $\lambda > 0$ small enough $\bar{x} + \lambda\hat{c}$ is in Ω° by property 5.1.9 of tangent cones. Therefore there exists a unique $\bar{\lambda} > 0$ be such that $\bar{x} + \bar{\lambda}\hat{c} \in \partial\Omega$ by strict convexity. We also have $l_{\bar{x}}(\hat{c}) = \bar{\lambda}$. Since the quantities involved in this theorem are invariant by isometry we can assume without loss of generality $T_{\bar{x}}\Omega = \{x \in \mathbb{R}^n \mid x_n = 0\}$ and $\bar{x} = 0$. In particular, our original hypotheses imply in this setting that there exists $f(x) : \bar{B}(0,r) \to [0,\varepsilon]$ such that

$$\partial\Omega \cap C^0_{r,\varepsilon} = \operatorname{graph}(f) \tag{2.4.59}$$

We now want to prove that

$$\frac{\delta}{\sqrt{1-\delta^2}} \leq rL_0h \tag{2.4.60}$$

implies $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon}$, or in other words that when (2.4.60) holds the second intersection between the ray $\{\lambda\hat{c} \mid \lambda \geq 0\}$ and $\partial\Omega$ is in $C^0_{r,\varepsilon}$. In order to show this it suffices to prove that the second intersection $\lambda_C\hat{c}$ of the ray $\{\lambda\hat{c} \mid \lambda \geq 0\}$ with $\partial C^0_{r,\varepsilon}$ is a point below graph$f$. Indeed we then have $\lambda\hat{c} \in \Omega^\circ \cap C^0_{r,\varepsilon}$ for λ small enough, $\lambda_C\hat{c} \in C^0_{r,\varepsilon} \setminus \Omega$ so that necessarily $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon} \cap \partial\Omega$ with $\bar{\lambda} \in (0, \lambda_C]$ by convexity.
Let $\hat{c} = (x_c, y_c) \in \mathbb{R}^{n-1} \times \mathbb{R}$ so that $\|x_c\| = \sqrt{1-\delta^2}$ and $y_c = \delta$. Then $\frac{r}{\sqrt{1-\delta^2}}\hat{c} = (x,y)$ with $\|x\| = r$, $y = r\frac{\delta}{\sqrt{1-\delta^2}}$, and in particular $f(x) \geq r^2L_0h \geq y$ by (2.4.60). Hence $(x,y) \in \partial C^0_{r,\varepsilon}$ lies below graph(f) and by the above reasoning $\bar{\lambda} \in (0, \frac{r}{\sqrt{1-\delta^2}}]$ with $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon}$.
We can now use this fact to prove the estimates (2.4.56), (2.4.57).
1. If $\bar{\lambda}\hat{c}$ is not in $C^0_{r,\varepsilon}$ then

$$\bar{\lambda} = \|\bar{\lambda}\hat{c}\| \geq \operatorname{dist}(0, \Omega \setminus C^0_{r,\varepsilon}) \geq \operatorname{dist}(0, T_\Omega(0) \setminus C^0_{r,\varepsilon}) \geq \min(r, \varepsilon) \tag{2.4.61}$$

It remains to consider the case $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon}$. In this setting we have by hypothesis $\partial\Omega \cap C^0_{r,\varepsilon} = \operatorname{graph}(f)$, so that $\bar{\lambda}\hat{c} \in \partial\Omega \cap C^0_{r,\varepsilon} = \operatorname{graph}(f)$. Then if $\bar{\lambda}\hat{c} = (y, f(y))$ we have

$$\frac{f(y)}{\|y\|} = \frac{\bar{\lambda}\|y_c\|}{\bar{\lambda}\|x_c\|} = \frac{\|y_c\|}{\|x_c\|} = \frac{\delta}{\sqrt{1-\delta^2}} \tag{2.4.62}$$

and since $f(y) \leq L_0\|y\|^2$ we deduce

$$\frac{\delta}{\sqrt{1-\delta^2}}\|y\| = f(y) \leq L_0\|y\|^2 \Rightarrow \frac{\delta}{\sqrt{1-\delta^2}L_0} \leq \|y\| \tag{2.4.63}$$

From this inequality together with (2.4.62) we can finally estimate

$$l_{\bar{x}}(\hat{c}) = \bar{\lambda} = \|\bar{\lambda}\hat{c}\| = \|(y, f(y))\| = (\|f(y)\|^2 + \|y\|^2)^{\frac{1}{2}} \geq \frac{\|y\|}{\sqrt{1-\delta^2}} \geq \frac{\delta}{L_0(1-\delta^2)} \tag{2.4.64}$$

Now if $\frac{\delta}{\sqrt{1-\delta^2}} \leq rL_0h$ we proved that $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon}$. Therefore

$$l^b_0(\delta) = \min\{l_0(\hat{c}) \mid \mathrm{dist}(\hat{c}, T_x\Omega) = \delta\} \geq \frac{\delta}{L_0(1-\delta^2)} \tag{2.4.65}$$

where we used (2.4.64) in the inequality. This proves the first case of (2.4.56).
In the other case, that is $\frac{\delta}{\sqrt{1-\delta^2}} > rL_0h$, if $\bar{\lambda}\hat{c}$ is not in $C^0_{r,\varepsilon}$ we have $l_0(\hat{c}) \geq \min(r, \varepsilon)$ by (2.4.61). If $\bar{\lambda}\hat{c}$ is in $C^0_{r,\varepsilon}$ then again $l_0(\hat{c}) \geq \frac{\delta}{L_0(1-\delta^2)}$. Putting together these two bounds we get the second part of (2.4.56)

$$l^b_0(\delta) = \min\{l_0(\hat{c}) \mid \mathrm{dist}(\hat{c}, T_x\Omega) = \delta\} \geq \min(\varepsilon, r, \frac{\delta}{L_0(1-\delta^2)}) \tag{2.4.66}$$

2. If $\bar{\lambda}\hat{c}$ is not in $C^0_{r,\varepsilon}$ we can give the very rough bound

$$l_0(\hat{c}) = \bar{\lambda} \leq D$$

Otherwise again $\bar{\lambda}\hat{c} = (y, f(y))$ with

$$\frac{\delta}{\sqrt{1-\delta^2}} = \frac{f(y)}{\|y\|} \geq \frac{L_0h\|y\|^2}{\|y\|} = L_0h\|y\| \tag{2.4.67}$$

which implies

$$l_0(\hat{c}) = \bar{\lambda} = \|(y, f(y))\| = (\|y\|^2 + \|f(y)\|^2)^{\frac{1}{2}} = (\|y\| + \frac{\delta^2}{1-\delta^2}\|y\|^2)^{\frac{1}{2}} \leq \frac{\delta}{L_0h(1-\delta^2)} \tag{2.4.68}$$

We can now prove the bound (2.4.57). If $\frac{\delta}{\sqrt{1-\delta^2}} \leq rL_0h$ then $\bar{\lambda}\hat{c} \in C^0_{r,\varepsilon}$ so that

$$l^B_0(\delta) = \min\{l_0(\hat{c}) \mid \mathrm{dist}(\hat{c}, T_0\Omega) = \delta\} \leq \frac{\delta}{L_0h(1-\delta^2)} \tag{2.4.69}$$

If $\frac{\delta}{\sqrt{1-\delta^2}} > rL_0h$ then the bound

$$l^B_0(\delta) \leq D \tag{2.4.70}$$

is trivial.
3. It is straightforward to check that the bounds given in point 1. and 2. are

increasing in δ. From this observation together with (2.4.56), (2.4.57) we get that if $\frac{\delta}{\sqrt{1-\delta^2}} > rL_0h$ or equivalently $\delta > s^{-1}(rL_0h) = \delta_*$ then

$$\begin{aligned} &\frac{l_0^b(k\delta,\delta)}{l_0^B(0,\delta)} = \frac{\min\{l_0(\hat{c}) \mid k\delta \leq \mathrm{dist}(\hat{c}, T_0\Omega) \leq \delta\}}{\max\{l_0(\hat{c}) \mid 0 \leq \mathrm{dist}(\hat{c}, T_0\Omega) \leq \delta\}} = \frac{\min\{l_0^b(\tau) \mid k\delta \leq \tau \leq \delta\}}{\max\{l_0^B(\tau) \mid 0 \leq \tau \leq \delta\}} \geq \\ &\geq \frac{\min(r, \varepsilon, k\delta/L_0(1-k^2\delta^2))}{D} \geq \frac{\min(r, \varepsilon, k\delta_*/L_0(1-k^2\delta_*^2))}{D} \end{aligned} \tag{2.4.71}$$

If $\delta \leq \delta_*$ then with the same reasoning

$$\begin{aligned} \frac{l_0^b(k\delta,\delta)}{l_0^B(0,\delta)} &\geq \frac{k\delta/(L_0(1-k^2\delta^2))}{\delta/(L_0h(1-\delta^2))} = \\ = \frac{kh(1-\delta^2)}{1-k^2\delta^2} &\geq \frac{kh(1-\delta)}{(1-k\delta)} \geq \frac{kh(1-\delta_*)}{(1-k\delta_*)} \end{aligned} \tag{2.4.72}$$

where in the last inequality we used that the term on the left is decreasing in δ and that $\delta \leq \delta_*$.
Recall that

$$R_{\Omega,\bar{x}}(k) = \inf_{0<\delta\leq 1} \frac{l_{\bar{x}}^b(k\delta,\delta)}{l_{\bar{x}}^B(0,\delta)} \tag{2.4.73}$$

and since (2.4.71) together with (2.4.72) give a lower bound on the ratio independent of δ and $\bar{x}$ also point 3 follows putting together these two results. □

This lemma allows us to easily prove the main theorem as a consequence of criterion 2.4.9.

Proof. We first show that for any fixed $k \in (0,1]$ there exists M_k such that

$$\inf_{\bar{x}\in\Omega} R_{\Omega,\bar{x}}(k) > M_k > 0 \tag{2.4.74}$$

By (2.4.58) we have $R_{\Omega,\bar{x}}(k) \geq \min_{1\leq i\leq 4}(a_i(\bar{x}))$ where $a_i(\bar{x})$ is the $i{-}th$ term appearing in the min argument of (2.4.58). Then $1 \leq i \leq 4$ it suffices to show that there exists a lower bound $M_k^i > 0$ for $a_i(\bar{x})$ independent from $\bar{x}$. This is obvious for $i = 1, 2$.
We now show that $L_{\bar{x}}$ is upper bounded. It follows immediately from the definitions $hL_{\bar{x}}r^2 \leq \varepsilon \leq D$: indeed in the setting introduced in Lemma 2.4.58, which is not restrictive with respect to the general case, we have $hL_{\bar{x}}r^2 \leq \min_{\|x\|=r} f(x) \leq \varepsilon$ and $\varepsilon \leq D$ otherwise the distance between $0 \in \Omega$ and the second intersection of $\{\lambda e_n \mid \lambda \geq 0\}$ with Ω would be greater than D. As a consequence, $rL_{\bar{x}}h$ is bounded so that $\delta_*(\bar{x}) = s^{-1}(rL_{\bar{x}}h) < m < 1$ for every $\bar{x}$ since $s : (0,1) \to \mathbb{R}_{>0}$ is a strictly increasing bijection. Hence

$$a_4(\bar{x}) = kh\frac{(1-\delta_*(\bar{x}))}{1-k\delta_*(\bar{x})} \geq kh(1-\delta_*(\bar{x})) \geq kh(1-m) = M_k^4$$

As for $i = 3$ we have

$$\begin{aligned} a_3(\bar{x}) = \frac{k\delta_*(\bar{x})}{DL\bar{x}(1-k^2\delta_*(\bar{x})^2)} &\geq \frac{k\delta_*(\bar{x})}{DL_{\bar{x}}} = \frac{krh\delta_*(\bar{x})}{s(\delta_*(\bar{x}))} = \\ = krh\sqrt{1-\delta_*(\bar{x})^2} &\geq krh\sqrt{1-m^2} = M_k^3 \end{aligned} \tag{2.4.75}$$

Since (2.4.74) holds with $M_k = \min_{1\leq i\leq 4}\{M_k^i\}$, we have all the hypotheses to apply Theorem (2.4.36) and conclude $\mathrm{NW}(\Omega) > 0$. □

Remark 2.4.15. We spend a few words to illustrate how the bound conditions are equivalent, under a few additional regularity assumptions, to the positive definiteness of the second fundamental form of Ω.
Let $\tilde{f} : \bar{B}(0, r) \to \mathbb{R}$ be the function whose graph locally represent $\partial\Omega$ with respect to a tangent plane in a fixed point $\bar{x}$.
When $\partial\Omega$ is a smooth submanifold of $\mathbb{R}^n$, $\tilde{f}$ is smooth, so the bound condition $hL_{\bar{x}}\|x\|^2 \leq f(x) \leq L\|x\|^2$ translate to a condition on the Taylor series of $\tilde{f}$. We will have, by the definition of tangent plane $\tilde{f}(x) = \frac{1}{2}x^T H(0)x + o(\|x\|^2)$ with $H(\bar{x})$ the Hessian of f. Then the bound conditions hold if and only if the Hessian $H(\bar{x})$ is positive definite in 0. If this condition holds for every $\bar{x} \in \partial\Omega$, then the second fundamental form of Ω is positive definite. However, it should be noted that our theorem requires also some uniformity condition on $o(x)$.

Remark 2.4.16. Under the same hypotheses an analogous lemma to 2.4.14 holds if $hL\|x\|^\alpha \leq f(x) \leq L\|x\|^\alpha$ with $\alpha > 1$. This seems to suggest that the bound we gave for NW(Ω) could be adapted also for sets with a few "singular points" where different bound conditions hold.
As an example, it is easy to check that the unit ball in every dimension respects the quadratic bound conditions uniformly, so that NW$(B(0, 1)) > 0$. By the theorem on polytopes, NW$(\Omega) > 0$ also for the ℓ_1 and ℓ_∞ balls. However, it is not yet clear whether NW$(\Omega) > 0$ for every ℓ_α ball, $\alpha \geq 1$.

2.5 Convergence theorem

In this section we analyze the convergence rate of a FDFW method using away directions as described in Table 2. In the appendix we explain how to extended the analysis for other variations.
Before stating the converge theorem we prove a straightforward lemma which will help us estimate the convergence rates:

Lemma 2.5.1. *Let $h_0 > 0$, $0 < \beta < 1$, $1 \leq r \leq 2$ and $\{h_k\}_{k\in\mathbb{N}}$ be a sequence satisfying*

$$h_k - h_{k+1} \geq \beta h_k^r \tag{2.5.1}$$

Then

$$h_k \leq \begin{cases} (1-\beta)^k h_0 & \text{if } r = 1 \\ (\frac{p}{\beta})^p k^{-p} & \text{if } 1 < r \leq 2 \end{cases} \tag{2.5.2}$$

where $p = \frac{1}{r-1}$.

Proof. In the case $r = 1$ inequality (2.5.1) can be rewritten as

$$h_{k+1} \leq (1-\beta)h_k \tag{2.5.3}$$

and the result follows by induction.
For $1 < r \leq 2$ let $H_r(x) = (\frac{p}{\beta})^p x^{-p}$ with $p = \frac{1}{r-1}$. We have for every $x > 0$

$$H_r'(x) = -p(\frac{p}{\beta})^p x^{-p-1} = -\beta(\frac{p}{\beta})^{p+1} x^{-p-1} = -\beta H_r(x)^r \tag{2.5.4}$$

By convexity

$$H_r(x+1) \geq H_r(x) + (x+1-x)H_r'(x) = H_r(x) + H_r'(x) \tag{2.5.5}$$

and applying (2.5.4) to the right hand side of (2.5.5) we get

$$H_r(x+1) \geq H_r(x) + H_r'(x) = H_r(x) - \beta H_r(x)^r \tag{2.5.6}$$

Let $k_0 = \frac{h_0}{\beta}^{-1/p}$ so that $H_r(k_0) = h_0$. We will now prove by induction

$$h_k \leq H_r(k_0 + k) \tag{2.5.7}$$

for every $k \in \mathbb{N}_0$. For $k = 0$ equation (2.5.7) follows by the definition of k_0. Assume we have $h_k \leq H_r(k+k_0)$ for some $k \in \mathbb{N}$, so that in particular $h_k = H_r(k+k_0+\Delta)$ for some $\Delta > 0$ since $H_r : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is decreasing and a bijection. Applying hypothesis (2.5.1) we get

$$h_{k+1} \leq h_k - \beta h_k^r = H_r(k + k_0 + \Delta) - \beta H_r(k + k_0 + \Delta)^r \tag{2.5.8}$$

and applying (2.5.6) to the RHS with $x = k_0 + k + \Delta$ we get

$$H_r(k + k_0 + \Delta) - \beta H_r(k + k_0 + \Delta)^r \leq H_r(k_0 + k + \Delta + 1) \leq H_r(k_0 + k + 1) \tag{2.5.9}$$

where the last inequality follows from the fact that H_r is decreasing. Concatenating (2.5.8) and (2.5.9) the inductive step is proved. We thus have for every $k \in \mathbb{N}$

$$h_k \leq H_r(k_0 + k) \leq H_r(k) = (\frac{p}{\beta})^p k^{-p} \tag{2.5.10}$$

□

We assume here that $f(x)$ has Lipschitz gradient

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ \forall \ x, y \in \Omega \tag{2.5.11}$$

and that it satisfies the Holder error bound condition

$$M(f(x) - f^*)^\theta \geq \mathrm{dist}(x, X) \ \forall \ x \in \Omega \setminus X \tag{2.5.12}$$

where $0 < \theta \leq 1$ and we recall that X is the set of minimizers for f.

Table 2: FDFW using the away directions
1. Let $x_0 \in \Omega$
2. **for** k = 0...T **do**
3. Let $s_k := \mathrm{LMO}_\Omega(\nabla f(x_k))$ and $d_k^{FW} := s_k - x_k$
4. Let $v_k := \mathrm{LMO}_{\mathcal{F}(x_k)}(-\nabla f(x_k))$ and $d_k^{\mathcal{A}} := x_k - v_k$
5. **if** $g_k^{FW} := (-\nabla f(x_k), d_k^{FW}) \leq \epsilon$ **then return** x_k
6. **if** $(-\nabla f(x_k), d_k^{FW})\|d_k^{\mathcal{A}}\| \geq (-\nabla f(x_k), d_k^{\mathcal{A}})\|d_k^{FW}\|$ **then**
7. $d_k := d_k^{FW}$ and $\alpha_{\max} := 1$
8. **else**
9. $d_k := d_k^{\mathcal{A}}$ and $\alpha_{\max} := \max\{\alpha \in \mathbb{R} \mid x_k + \alpha d_k^{\mathcal{A}} \in \Omega\}$
10. **end if**
11. $\alpha_k = \min(\frac{(-\nabla f(x_k), d_k)}{\|d_k\|^2 L}, \alpha_{\max})$ *minimize the upper bound of f*
12. Update $x_{k+1} := x_k + \alpha_k d_k$
13. **end for**

By Proposition 2.3.6 equation (2.5.12) implies

$$\|\pi(T_\Omega(x), -\nabla f(x))\| \geq \frac{(f(x) - f^*)^{1-\theta}}{M} \ \forall \ x \in \Omega \setminus X \tag{2.5.13}$$

In our convergence rate estimates we will use only (2.5.13) and not (2.5.12) directly, so that we could only assume (2.5.13) and prove the same convergence properties. However, for $\theta > 0$

$$\frac{M}{\theta}(f(x) - f^*)^\theta \geq \mathrm{dist}(x, X) \ \forall \ x \in \Omega \setminus X \tag{2.5.14}$$

whenever (2.5.13) holds, so that the error bound conditions are actually necessary in our convergence proof, up to the constant θ. For $\theta = 0$ the error bound condition is trivially satisfied by every continuous function f with $M = \mathrm{diam}(\Omega)$.
If f is μ - strongly convex than for every $\bar{x} \in X$, $x \in \Omega$

$$f(x) - f^* = f(x) - f(\bar{x}) \geq f(x) + (x - \bar{x}, \nabla f(\bar{x})) + \frac{\mu}{2}\|x - \bar{x}\|^2 \geq \frac{\mu}{2}\|x - \bar{x}\|^2 \tag{2.5.15}$$

were in the last inequality we used $(x - \bar{x}, \nabla f(\bar{x})) \geq 0$ by first order optimality conditions. We can rewrite (2.5.15) as

$$\sqrt{\frac{2}{\mu}}(f(x) - f^*)^{\frac{1}{2}} \geq \|x - \bar{x}\| \geq \mathrm{dist}(x, X) \tag{2.5.16}$$

so that we retrieve the Holder error bound condition with $\theta = 1/2$ and $M = (\frac{2}{\mu})^{\frac{1}{2}}$. In the statement of the convergence theorem we will use

$$\dim_2(\Omega) = \max_{\substack{\mathcal{F} \text{ is a face of } \Omega \\ \mathcal{F} \subsetneq \Omega}} \dim(\mathcal{F})$$

Theorem 2.5.2. *Assume that f satisfies the error bound condition* (2.5.12)*, that it has $L-$ Lipschitz gradient and that* $\mathrm{NW}^f(\Omega) > 0$*. Then for the sequence generated by the FDFW in Table 2:*

$$f(x_k) - f^* \leq \begin{cases} \max(\frac{h_0}{2^{q(k)}}, (\frac{2pLM^2}{\mathrm{NW}^f(\Omega)^2})^p q(k)^{-p}) & \text{for } 0 \leq \theta < \frac{1}{2} \\ \max(\frac{1}{2}, (1 - \frac{\mathrm{NW}^f(\Omega)^2}{2M^2L}))^{2q(k)} h_0 & \text{for } \theta = \frac{1}{2} \end{cases} \tag{2.5.17}$$

for $p = \frac{1}{1-2\theta}$ *and* $q(k) = \lfloor \frac{k}{2(\dim_2(\Omega)+2)} \rfloor$*. If* $\frac{1}{2} < \theta \leq 1$ *then* $h_k = O(\frac{1}{2^{q(k)}})$*.*

The proof roughly follows the same steps of [39], Theorem 4 which proves linear convergence for the AFW method on polytopes in the strongly convex case. We will study how the decrease of the objective function $f(x_k) - f(x_{k+1})$ relates to the current gap $f(x_k) - f^*$ in 3 possible cases. When $\alpha_k < \alpha_{\max}$ the decrease can be related to the square of the slope of $-\nabla f(x_k)$ along the descent direction d_k. In turns this slope can be related to the current gap using the error bound condition and the hypothesis $\mathrm{NW}^f(\Omega) > 0$. When $d_k = d_k^{FW}$ and $\alpha_k = \alpha_{\max} = 1$ the Lipschitz condition on the gradient alone implies $f(x_k) - f(x_{k+1}) \geq \frac{1}{2}(f(x_k) - f^*)$. Finally, when the algorithm makes a maximal away step we cannot control how much f decreases because we can't give a lower bound on the size of the step. However, since under these conditions $\mathcal{F}(x_{k+1}) < \mathcal{F}(x_k)$ we can bound the number of maximal away steps as a fraction of the total steps.

Proof. Let $h_k = f(x_k) - f^*$, $r_k = -\nabla f(x_k)$ and $p_k = \|\pi(T_\Omega(x_k), -\nabla f(x_k))\|$. Let

$$\begin{aligned} m_k &= \text{argmin}\{(\nabla f(x_k), v) \mid v \in \Omega\} \\ M_k &= \text{argmax}\{(\nabla f(x_k), s) \mid s \in \mathcal{F}(x_k)\} \end{aligned} \tag{2.5.18}$$

so that for the s_k and v_k selected by the minimization oracle we have $s_k \in m_k$ and $v_k \in M_k$ respectively. We divide our analysis in 3 cases, and for $i = 1, 2, 3$ introduce the functions $n_i(k)$ giving the number of times that case i occurs in the first k steps. We will assume from now on $x_k \notin m_k$, otherwise $g_k^{FW} = 0$ and the algorithm stops with $f(x_k) = f^*$ by first order optimality conditions.

Case 1: $\alpha_k < \alpha_{\max}$. Then the error bound condition (2.5.12) implies condition (2.5.13) so that

$$h_k^{1-\theta} \leq M p_k \tag{2.5.19}$$

Moreover by the standard descent lemma (see [6], proposition 6.1.2)

$$f(x_k + \alpha d_k) \leq f(x_k) + \alpha(\nabla f(x_k), d_k) + \frac{\alpha^2 L}{2}\|d_k\|^2 \tag{2.5.20}$$

Minimizing the right hand side with respect to α we have that for $\alpha = \alpha_k$

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \leq f(x_k) - \frac{1}{2L\|d_k\|^2}(r_k, d_k)^2 \tag{2.5.21}$$

which bringing $f(x_k)$ on the left hand side becomes

$$h_k - h_{k+1} \geq \frac{1}{2L\|d_k\|^2}(r_k, d_k)^2 = \frac{1}{2L}(r_k, \frac{d_k}{\|d_k\|})^2 \tag{2.5.22}$$

where we used $h_k - h_{k+1} = (f(x_k) - f^*) - (f(x_{k+1}) - f^*) = f(x_k) - f(x_{k+1})$. Now by definition

$$\begin{aligned} \text{dirNW}(\Omega, x_k, r_k) &= \inf_{\substack{s \in m_k \\ v \in M_k}} \frac{(s - v, r_k)}{(\|v - x_k\| + \|s - x_k\|)p_k} = \inf_{\substack{s \in m_k \\ v \in M_k}} \frac{(s - x_k - (v - x_k), r_k)}{(\|v - x_k\| + \|s - x_k\|)p_k} \leq \\ &\leq \frac{(s_k - x_k - (x_k - v_k), r_k)}{p_k(\|v_k - x_k\| + \|s_k - x_k\|)} \end{aligned} \tag{2.5.23}$$

We now distinguish two cases. If $d_k^A = 0$ then $d_k = d_k^{FW} = s_k - x_k$ and

$$\frac{(s_k - x_k - (v_k - x_k), r_k)}{p_k(\|v_k - x_k\| + \|s_k - x_k\|)} = \frac{(s_k - x_k, r_k)}{p_k\|s_k - v_k\|} = \frac{(d_k, r_k)}{p_k\|d_k\|} \tag{2.5.24}$$

If $d_k^A \neq 0$ then

$$(r_k, \frac{d_k}{\|d_k\|}) = \max((r_k, \frac{d_k^{FW}}{\|d_k^{FW}\|}), (r_k, \frac{d_k^A}{\|d_k^A\|})) \tag{2.5.25}$$

and

$$\begin{aligned} &\frac{1}{p_k}\frac{((s_k - x_k) + (x_k - v_k), r_k)}{\|s_k - x_k\| + \|v_k - x_k\|} = \frac{1}{p_k}\frac{(d_k^A + d_k^{FW}, r_k)}{\|d_k^A\| + \|d_k^{FW}\|} \leq \\ \leq &\frac{1}{p_k}\max((r_k, \frac{d_k^{FW}}{\|d_k^{FW}\|}), (r_k, \frac{d_k^A}{\|d_k^A\|})) = \frac{1}{p_k}(r_k, \frac{d_k}{\|d_k\|}) \end{aligned} \tag{2.5.26}$$

Now the left hand side of equations (2.5.24) and (2.5.26) is also an upper bound for $\mathrm{dirNW}(\Omega, x_k, r_k)$ by (2.5.23). Hence in both cases we have

$$\mathrm{dirNW}(\Omega, x_k, r_k) \leq \frac{1}{p_k}(r_k, \frac{d_k}{\|d_k\|}) \tag{2.5.27}$$

and isolating the scalar product

$$p_k \mathrm{dirNW}(\Omega, x_k, r_k) \leq (r_k, \frac{d_k}{\|d_k\|}) \tag{2.5.28}$$

Since by definition $\mathrm{NW}^f(\Omega) \leq \mathrm{dirNW}(\Omega, x_k, r_k)$

$$p_k \mathrm{NW}^f(\Omega) \leq (r_k, \frac{d_k}{\|d_k\|}) \tag{2.5.29}$$

We finally have

$$h_k - h_{k+1} \geq \frac{1}{2L}(r_k, \frac{d_k}{\|d_k\|})^2 \geq \frac{1}{2L} p_k^2 \mathrm{NW}^f(\Omega)^2 \geq \frac{\mathrm{NW}^f(\Omega)^2}{2LM^2} h_k^{2-2\theta} \tag{2.5.30}$$

were we used (2.5.22) in the first inequality, (2.5.29) in the second inequality and (2.5.19) in the third one respectively.

Case 2: $\alpha_k = \alpha_{\max} = 1, d_k = d_k^{FW}$. First, notice that for any $x^* \in X$

$$h_k = f(x_k) - f(x^*) \leq (-\nabla f(x_k), x^* - x_k) \tag{2.5.31}$$

Now since by hypothesis $d_k = d_k^{FW} = s_k - x_k$ with $s_k \in m_k = \mathrm{argmin}_{x \in \Omega}(\nabla f(x_k), x)$ we have $(\nabla f(x_k), x^*) \geq (\nabla f(x_k), s_k)$ so that

$$(\nabla f(x_k), x_k - x^*) \leq (\nabla f(x_k), x_k - s_k) = (-\nabla f(x_k), d_k)$$

Combining this equation with (2.5.31) we obtain

$$h_k \leq (-\nabla f(x_k), d_k) \tag{2.5.32}$$

Again by the standard descent lemma applied to f with center x_k and $\alpha = 1$

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

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

$$\frac{(-\nabla f(x_k), d_k)}{\|d_k\|^2 L} \geq 1 \Rightarrow -L\|d_k\|^2 \geq (\nabla f(x_k), d_k) \tag{2.5.33}$$

so that

$$h_k - h_{k+1} \geq f(x_k) - f(x_{k+1}) \geq (-\nabla f(x_k), d_k) + \frac{L}{2}\|d_k\|^2 \geq -\frac{1}{2}(\nabla f(x_k), d_k) \tag{2.5.34}$$

Concatenating (2.5.32) multiplied by 1/2 to this inequality we get $h_k - h_{k+1} \geq \frac{1}{2} h_k$.

Case 3: $\alpha_k = \alpha_{\max}$, $d_k = d_k^{\mathcal{A}}$. Let $D_\Omega(\bar{x}) = \dim(\mathcal{F}(\bar{x}))$. Then since $x_{k+1} = x_k + d_k$ lies on the boundary of $\mathcal{F}(x_k)$, we have $D_\Omega(x_{k+1}) \leq D_\Omega(x_k) - 1$. Thanks to this relation it is easy to see that there can not be $\dim_2(\Omega) + 2$ consecutive case 3 steps. Indeed, an easy induction shows that if the algorithm does j case 3 consecutive steps from the iteration k to the iteration $k + j - 1$, then it produces a center x_{k+j} such that $\dim(\mathcal{F}(x_{k+j})) \leq \dim_2(\Omega) - j + 1$.

The analysis of case 3 proves that among $\dim_2(\Omega) + 2$ consecutive steps there must be at least one not in case 3, so that necessarily

$$n_1(k) + n_2(k) \geq \lfloor \frac{k}{\dim_2(\Omega) + 2} \rfloor \tag{2.5.35}$$

In particular

$$\max(n_1(k), n_2(k)) \geq \frac{1}{2} \lfloor \frac{k}{\dim_2(\Omega) + 2} \rfloor \geq q(k) \tag{2.5.36}$$

Let $i(n)$ be the $n-$th index for which the step size is not maximal, or in other words respecting the conditions of case 1. Notice that $i(\cdot)$ is defined on a subset I of $\mathbb{N}$ such that $|I|$ is the number of case 1 steps.
Using $i(n)$ we can rewrite the number of case 1 step in the first k iterations as

$$n_1(k) = \max(n \in \mathbb{N}_0 \mid i(n) < k) \tag{2.5.37}$$

Then

$$\begin{aligned} i(n_1(k)) &= i(\max(n \in \mathbb{N}_0 \mid i(n) < k)) < k \\ i(n_1(k) + 1) &= i(\max(n \in \mathbb{N}_0 \mid i(n) < k) + 1) \geq k \end{aligned} \tag{2.5.38}$$

We define $j(n)$ analogously for case 2, so that

$$\begin{aligned} j(n_2(k)) &= j(\max(n \in \mathbb{N}_0 \mid j(n) < k)) < k \\ j(n_2(k) + 1) &= j(\max(n \in \mathbb{N}_0 \mid j(n) < k) + 1) \geq k \end{aligned} \tag{2.5.39}$$

By (2.5.36) at least one between $n_1(k)$ and $n_2(k)$ is greater than or equal to $q(k)$. If $n_1(k) \geq q(k)$ then $i(q(k)) \leq i(n_1(k)) < k$, so that by monotonicity $h_k \leq h_{i(q(k))}$. Analogously if $n_2(k) \geq q(k)$ then $h_k \leq h(j(q(k)))$. Summarizing

$$h_k \leq \max(h_{i(q_k)}, h_{j(q(k))}) \tag{2.5.40}$$

We now examine what happens for θ varying in $[0, 1)$.

$\boldsymbol{0 \leq \theta < \frac{1}{2}}$. Then the sequence $\{l_n\}_{n \in I} = \{h_{i(n)}\}_{n \in I}$ satisfies the hypotheses of Lemma 2.5.1 with $r = 2 - 2\theta$ and $\beta = \frac{\mathrm{NW}^f(\Omega)^2}{2LM^2}$. Indeed

$$l_{n+1} = h_{i(n+1)} \leq h_{i(n)+1} \leq h_{i(n)} - \frac{\mathrm{NW}^f(\Omega)^2}{2LM^2} h_{i(n)}^{2-2\theta} = l_n - \frac{\mathrm{NW}^f(\Omega)^2}{2LM^2} l_n^{2-2\theta} \tag{2.5.41}$$

by equation (2.5.30). Hence by Lemma 2.5.1

$$h_{i(q_k)} = l_{q(k)} \leq (\frac{2pLM^2}{N^f(\Omega)^2})^p q(k)^{-p} \tag{2.5.42}$$

with $p = \frac{1}{1-2\theta}$. Analogously since $h_{j(q(k+1))} \leq \frac{1}{2} h_{j(q(k))}$ we have

$$h_{j(q(k))} \leq \frac{1}{2^{q(k)}} h_0 \tag{2.5.43}$$

Now using (4.6.3) and (4.6.10) to bound the right hand side of (2.5.40) we get exactly the thesis for $\theta \in [0, \frac{1}{2})$.

$\boldsymbol{\theta = \frac{1}{2}}$. Let C_i be the set of indexes for which the method does a case i step, for $i = 1, 2$. For every $n \in C_1$ we have

$$h_{n+1} \leq (1 - \frac{\mathrm{NW}^f(\Omega)^2}{2M^2 L}) h_n \tag{2.5.44}$$

and for every $n \in C_2$

$$h_{n+1} \leq \frac{1}{2} h_n \tag{2.5.45}$$

Since equations (2.5.44) and (2.5.45) hold for $n_1(k)$ and $n_2(k)$ distinct values of n smaller than k respectively we get by induction

$$\begin{aligned} h_k &\leq h_0 (1 - \frac{\mathrm{NW}^f(\Omega)}{2M^2 L})^{n_1(k)} (\frac{1}{2})^{n_2(k)} \leq h_0 \max(\frac{1}{2}, (1 - \frac{\mathrm{NW}^f(\Omega)}{2M^2 L}))^{n_1(k)+n_2(k)} \leq \\ &\leq h_0 \max(\frac{1}{2}, (1 - \frac{\mathrm{NW}^f(\Omega)}{2M^2 L}))^{2q(k)} \end{aligned} \tag{2.5.46}$$

$\boldsymbol{\frac{1}{2} < \theta < 1}$. Let $\beta = \frac{\mathrm{NW}^f(\Omega)^2}{2LM^2}$ and $r = 2 - 2\theta$. Exactly as in the case $0 \leq \theta < \frac{1}{2}$ we get equation (2.5.41)

$$l_{n+1} \leq l_n - \beta l_n^r \tag{2.5.47}$$

We want to show that the sequence $\{l_n\}_{n \in I}$ has at most $(N - f^*)/(\beta^{1/(1-r)}) + 1$ terms, where $N = \sup_{x \in \Omega} f(x)$. Since $r < 1$ we have $l_n - \beta l_n^r \leq 0$ for $l_n \leq \beta^{1/(1-r)}$.
For every $n \in I$ different from $\sup I$ we have $l_n > \beta^{1/(1-r)}$, otherwise l_{n+1} would be ≤ 0, contradicting the strict positivity of the sequence. Thus

$$l_{n+1} \leq l_n - \beta l_n^r \leq l_n - \beta^{\frac{1}{1-r}} \tag{2.5.48}$$

for every $n \in I \setminus \sup I$. By induction

$$0 \leq l_{n+1} \leq l_1 - n\beta^{\frac{1}{1-r}} \leq N - f^* - n\beta^{\frac{1}{1-r}} \tag{2.5.49}$$

where we used $l_1 \leq h_0 = f(x_0) - f^* \leq N - f^*$. From (2.5.49) it follows immediately

$$n \leq \frac{N - f^*}{\beta^{\frac{1}{1-r}}}$$

and the uniform bound on the length of $\{l_n\}_{n \in I}$ is proved.
We therefore have

$$n_1(k) \leq \frac{N - f^*}{\beta^{\frac{1}{1-r}}} = N_{f,\theta} \tag{2.5.50}$$

for every $k \in \mathbb{N}$ so that

$$n_2(k) \geq 2q(k) - N_{f,\theta} \tag{2.5.51}$$

by (2.5.35) and consequently

$$h_k \leq \frac{h_0}{2^{n_2(k)}} \leq \frac{h_0 2^{N_{f,\theta}}}{2^{2q(k)}} = O(\frac{1}{2^{2q(k)}}) \tag{2.5.52}$$

□

When we start the algorithm from a vertex of Ω, we get the following result:

Corollary 2.5.3. *If the algorithm in Table 2 starts from a vertex of Ω, then the same results hold with*

$$q(k) = \frac{k}{2(\dim_2(\Omega) + 2)}$$

Proof. We will use the notation introduced in the proof of the main theorem. Since the algorithm start from a vertex the first step must in the FW direction d_0^{FW}, and in particular not a case 3 step. Since there can be at most $\dim_2(\Omega)+1$ case 3 consecutive steps.

$$n_1(k) + n_2(k) \geq 1 + \lfloor\frac{k-1}{\dim_2(\Omega)+2}\rfloor \geq \frac{k}{\dim_2(\Omega)+2}$$

The conclusion follows as in the main theorem. □

If Ω is strictly convex then $\dim_2(\Omega) = 0$ so we get the following:

Corollary 2.5.4. *Assume Ω is strictly convex. Then the results of the main theorem hold with $q(k) = \lfloor k/4 \rfloor$.*

For the simplex Δ^n the estimate can be improved with an argument equivalent to the one used originally in [34] for the AFW method: it is easy to see that $\dim(\mathcal{F}(x_{k+1})) \leq \dim(\mathcal{F}(x_k)) + 1$ in case 1 and 2. As a consequence, whenever the algorithm starts from a vertex $\dim \mathcal{F}(x_k) = n_1(k) + n_2(k) - n_3(k)$ so that $n_1(k) + n_2(k) \geq n_3(k)$ and the corollary below follows:

Corollary 2.5.5. *Assume $\Omega = \Delta_n$ and that algorithm 1 start from a vertex. Then the same results of the main theorem hold with $q(k) = k/2$.*

2.5.1 Inexact oracles

. When dealing with inexact oracles the main obstacle to generalize Theorem 2.5.2 is that the approximated solution of the linear subproblem may be far from the actual set of minimizers. As a consequence the slope along the approximated search direction may be arbitrarily smaller than the slope along the actual search direction, even for solvers with small error on the objective. Using more explicit equations, if $\tilde{d}_k$ is the approximated search direction then $(-\nabla f(x_k), \tilde{d}_k) \approx (-\nabla f(x_k), \tilde{d}_k)$ does not imply in general $\tilde{d}_k \approx d_k$. Since ensuring that $(-\nabla f(x_k), \frac{\tilde{d}_k}{\|\tilde{d}_k\|})$ is large enough is fundamental in the proof of Theorem (2.5.2), it becomes necessary to add the assumption that $x_k + \tilde{d}_k$ is close enough to the set of minimizers as a property of the oracle. For this reason we will use an approximating oracle $\mathrm{LMO}_C(r, \delta, \mu) \in C$ such that

$$(r, \mathrm{LMO}_C(r, \delta, \mu)) - \min\{(r, x) \mid x \in \Omega\} \leq \delta \tag{2.5.53}$$

and

$$\mathrm{dist}(\mathrm{LMO}_C(r, \delta, \mu), \mathrm{argmin}\{(r, x) \mid x \in \Omega\}) \leq \mu \tag{2.5.54}$$

for every $r \in \mathbb{R}^n$, $\delta, \mu > 0$ and C varying among the faces of Ω. Before stating the convergence theorem we show a couple of examples when this oracle may be possible to build starting from "simpler" oracles.

Example 1. Assume that Ω is a polytope represented as $\{x \in \mathbb{R}^n \mid Ax \leq b\}$, and that the linear minimization oracle can correctly identify the set of active constraints I related to the problem $\min\{(r, x) \mid x \in \mathcal{F}\}$ for every face $\mathcal{F}$ of Ω. Assume also that the error in solving the corresponding linear system $A_I x = b_I$ is at most ϵ, where ϵ could be some function of the machine precision and the dimension of the problem. In other words assume that for every possible I representing a vertex we have $\|\tilde{x} - A_I^{-1} b_I\| \leq \epsilon$, where $\tilde{x}$ is the solution computed by the minimization oracle. Then the error on the objective is at most $(r, \tilde{x}) - (r, A_I^{-1} b_I) \leq \|r\|\epsilon$. If $r = \pm\nabla f(x)$ then $\|r\|\epsilon = \|\nabla f(x)\|\epsilon \leq M\epsilon$ with $M = \max_{x \in \Omega} \|\nabla f(x)\|$. In conclusion, if the approximated solutions of the linear system of active constraints are in Ω we have an oracle $\mathrm{LMO}_\Omega(r, \delta, \mu)$ for every $\mu \leq \epsilon$ and $\delta \leq M\epsilon$.

Example 2. Assume that Ω is a strictly convex set, and assume that we have an upper bound oracle

$$\mathrm{LMO}_\Omega^u(r, \delta) \in \Omega, \tag{2.5.55}$$

Then (see Proposition 5.5.1 of the appendix) there exists $\delta > 0$ and a strictly increasing function $m_\Omega : [0, \tau] \to \mathbb{R}_{\geq 0}$ such that $m_\Omega(0) = 0$ and

$$m_\Omega(\delta) \geq \max_{r \in \mathbb{R}^n \setminus \{0\}} \mathrm{dist}(\mathrm{LMO}_\Omega^u(r, \delta\|r\|), \mathrm{argmin}_{x \in \Omega}(r, x)) \tag{2.5.56}$$

for every $\delta \in [0, \tau]$.

Table 3: FDFW with approximated oracle
1. Let $x_0 \in \Omega$, $(\delta_0, \mu_0) \in (0, 1)^2$, $D > 0$
2. **for** k = 0...T **do**
3. **if** $(\nabla f(x_k), \mathrm{LMO}_\Omega(r, \epsilon, D)) \geq (\nabla f(x_k), x_k)$ **then return** x_k
4. Set $(\tilde{d}_k^{\mathrm{FW}}, \delta_{k+1}, \mu_{k+1}) = \mathrm{ApDirections}(\nabla f(x_k), \delta_k, \mu_k, x_k, \Omega)$
5. **if** $(-\nabla f(x_k), \mathrm{LMO}_{\mathcal{F}(x_k)}(-\nabla f(x_k), \epsilon, D)) \geq (-\nabla f(x_k), x_k)$ **then**
6. Set $(-\tilde{d}_k^{\mathrm{A}}, \delta_0, \mu_0) = \mathrm{ApDirections}(-\nabla f(x_k), \delta_0, \mu_0, x_k, \mathcal{F}(x_k))$
7. **else:** go to step 10
8. **end if**
9. **if** $(-\nabla f(x_k), \frac{\tilde{d}_k^{FW}}{\|\tilde{d}_k^{FW}\|}) \geq (-\nabla f(x_k), \frac{\tilde{d}_k^{\mathcal{A}}}{\|\tilde{d}_k^{\mathcal{A}}\|})$ **then**
10. $\tilde{d}_k := \tilde{d}_k^{FW}$ and $\alpha_{\max} := 1$
11. **else**
13. $\tilde{d}_k := \tilde{d}_k^{\mathcal{A}}$ and $\alpha_{\max} := \max\{\alpha \in \mathbb{R} \mid x_k + \alpha\tilde{d}_k^{\mathcal{A}} \in \Omega\}$
14. **end if**
15. $\alpha_k = \min(\frac{(-\nabla f(x_k), \tilde{d}_k)}{\|\tilde{d}_k\|^2 L}, \alpha_{\max})$
16. Update $x_{k+1} := x_k + \alpha_k \tilde{d}_k$
17. **end for**

Table 4: procedure ApDirections (r, δ, μ, x, C)
1. Let $d := 0$, $(\rho, \beta) \in (0,1)^2$.
2. **if** $\delta < \beta(d, r)$ **and** $\mu < \beta\|d\|$ **return** (d, δ, μ)
3. **else:**
4. $\delta := \rho\delta$, $\mu := \rho\mu$
5. $d := \mathrm{LMO}_C(r, \delta, \mu) - x$
6. Go to step 2.

This function m_Ω gives an upper bound on the distance of $\mathrm{LMO}^u_\Omega(r, \delta)$ from the set of actual minimizers given the approximation error on the objective. We then have an oracle

$$\mathrm{LMO}^u_\Omega(r, \delta, \mu) = \mathrm{LMO}^u_\Omega(r, \delta)$$

for every $\mu \geq m(\frac{\delta}{\|r\|})$, since by equation (2.5.56)

$$\mathrm{dist}(\mathrm{LMO}^u_\Omega(r, \delta), \mathrm{argmin}_{x \in \Omega}(r, x)) \leq m(\frac{\delta}{\|r\|}) \tag{2.5.57}$$

In the algorithm described in Table 2 we use the procedure ApDirections which guarantees

$$(-\nabla f(x_k), \frac{\tilde{d}_k}{\|\tilde{d}_k\|}) \geq \bar{\beta}(-\nabla f(x_k), \frac{d_k}{\|d_k\|}) \tag{2.5.58}$$

whenever both the classical and the away FW directions are computed with $\bar{\beta} \in (0, 1)$ some function of the algorithm's parameters. Notice that with respect to Table 2 we anticipated the stopping criterion to step 3 and also inserted a preliminary condition before computing the away step. These modifications are necessary to ensure finite termination of the procedure ApDirections.

Lemma 2.5.6. *If $\epsilon > 0$, the procedure ApDirections terminates in a finite number of iterations when called by the algorithm in Table 3.*

Proof. We distinguish two cases. If $(\nabla f(x_k), x_k) - \min_{y \in \Omega}(\nabla f(x_k), y) \leq \epsilon$ then the algorithm in Table 3 returns x_k at step 3 and does not call Apdirections. Otherwise, let i be the number of cycles performed by an ApDirections instance called in step 4, and let δ'_i, d'_i, μ'_i be the values of δ, d, μ in cycle i. Then we clearly have $\delta'_i = \rho^i \delta'_0 \to 0$ for $i \to \infty$. But then since $r = \nabla f(x_k)$

$$\begin{aligned} \beta(d'_i, r) = (\mathrm{LMO}_C(\nabla f(x_k), \delta'_i, \mu'_i) - x_k, \nabla f(x_k)) \geq \\ \geq (\nabla f(x_k), x_k) - \min_{y \in \Omega}(\nabla f(x_k), y) - \delta'_i \geq \varepsilon - \delta'_i > \delta'_i \end{aligned} \tag{2.5.59}$$

where the last inequality holds for i large enough. The analysis for the instances called by step 6 is completely analogous. □

We can now prove a converge theorem analogous to the one proved for exact oracles:

Theorem 2.5.7. *Assume that in the algorithm of Table 3 $\epsilon > 0$. Then under the hypotheses of Theorem 2.5.2 the algorithm stops after a finite number of iterations M, and for $0 \leq \theta \leq \frac{1}{2}$ it has the same convergence rate with*

$$\tilde{\mathrm{NW}}^f(\Omega) = \frac{1-\beta}{2+2\beta}\mathrm{NW}^f(\Omega) \tag{2.5.60}$$

instead of $\mathrm{NW}^f(\Omega)$ *and* $\frac{1+2\beta}{2(1+\beta)}$ *instead of* $\frac{1}{2}$ *as base of the exponential term.*

Proof. First we have by Lemma 2.5.6 that the procedure ApDirections always terminate in a finite number of iterations. We now claim that

$$(-\nabla f(x_k), \tilde{d}_k^{FW}) \geq \frac{(-\nabla f(x_k), d_k^{FW})}{1+\beta} \tag{2.5.61}$$

Indeed the return condition in the procedure ApDirections dictates

$$\delta_{k+1} \leq \beta(\tilde{d}_k^{FW}, -\nabla f(x_k)) \tag{2.5.62}$$

with

$$\tilde{d}_k^{FW} = \mathrm{LMO}_C(\nabla f(x_k), \delta_{k+1}, \mu_{k+1}) \tag{2.5.63}$$

But then

$$\begin{aligned}(-\nabla f(x_k), d_k^{FW}) &= (\nabla f(x_k), x_k) - \min\{(\nabla f(x_k), y) \mid y \in \Omega\} \leq \\ &\leq (\nabla f(x_k), x_k - \mathrm{LMO}_C(\nabla f(x_k), \delta_{k+1}, \mu_{k+1}, \Omega)) + \delta_{k+1} \leq \\ &\leq (\nabla f(x_k), x_k - \mathrm{LMO}_C(\nabla f(x_k), \delta_{k+1}, \mu_{k+1}, \Omega)) + \beta(-\nabla f(x_k), \tilde{d}_k) = \\ &= (1+\beta)(-\nabla f(x_k), \tilde{d}_k)\end{aligned} \tag{2.5.64}$$

where we applied (2.5.62) and (2.5.63) in the last inequality and in the last equality respectively.
With the same proof We also have

$$(-\nabla f(x_k), \tilde{d}_k^{AW}) \geq \frac{(-\nabla f(x_k), d_k^{AW})}{1+\beta} \tag{2.5.65}$$

whenever the condition in step 5 is satisfied.
As for d_k^{FW}, the return condition on Apdirections implies that there exists $y_k \in \mathrm{argmin}\{y \in \Omega \mid (\nabla f(x_k), y)\}$ such that

$$\|x_k + \tilde{d}_k^{FW} - y_k\| < \beta\|\tilde{d}_k^{FW}\| \tag{2.5.66}$$

We can assume $d_k^{FW} = y_k - x_k$ since the proof of Theorem 2.5.2 does not depend on a particular oracle. Then equation (2.5.66) can be rewritten as

$$\|\tilde{d}_k^{FW} - d_k^{FW}\| < \beta\|d_k^{FW}\| \tag{2.5.67}$$

which implies

$$\frac{1}{\|\tilde{d}_k^{FW}\|} \geq \frac{1-\beta}{\|d_k^{FW}\|} \tag{2.5.68}$$

Again with the same proof

$$\frac{1}{\|\tilde{d}_k^{AW}\|} \geq \frac{1-\beta}{\|d_k^{AW}\|} \tag{2.5.69}$$

whenever the condition in step 5 is satisfied. To conclude we analyze a few details about how the analysis of the 3 cases done in Theorem (2.5.2) adapts for this approximated algorithm.

Case 1: $\alpha_k < \alpha_{\max}$. Multiplying (2.5.61) by (2.5.68) we get

$$(-\nabla f(x_k), \frac{\tilde{d}_k^{FW}}{\|\tilde{d}_k^{FW}\|}) \geq \frac{1-\beta}{1+\beta}(-\nabla f(x_k), \frac{d_k^{FW}}{\|d_k^{FW}\|}) \tag{2.5.70}$$

Whenever the condition in step 5 is satisfied reasoning analogously on $\tilde{d}_k^{AW}$ and passing to the max we get the same inequality relating $\tilde{d}_k$ with d_k. Then by equation (2.5.28) (we recall that $r_k = -\nabla f(x_k)$):

$$(r_k, \frac{\tilde{d}_k}{\|\tilde{d}_k\|}) \geq \frac{1-\beta}{1+\beta}(r_k, \frac{d_k}{\|d_k\|}) \geq \frac{1-\beta}{1+\beta} p_k \mathrm{NW}^f(\Omega) \tag{2.5.71}$$

It remains to analyze what happens when the condition in step 5 is not satisfied. In this setting we have $(-\nabla f(x_k), d_k^{FW}) \geq \epsilon$ and $(-\nabla f(x_k), d_k^{AW}) < \epsilon$. Then

$$(r_k, d_k^{FW} + d_k^{AW}) \leq 2(r_k, d_k^{FW})$$

and we can apply this to bound $(\tilde{d}_k^{FW}, r_k)$ in terms of $\mathrm{NW}^f(\Omega)$ as we've already done for the exact algorithm

$$\begin{aligned}
&\mathrm{NW}^f(\Omega) \leq \mathrm{dirNW}(\Omega, x_k, r_k) = \frac{(r_k, q^*(r_k) - s^*(r_k))}{\|p_k\|(\|q^*(r_k) - x_k\| + \|s^*(r_k) - x_k\|)} = \\
=&\frac{(r_k, d_k^{AW} + d_k^{FW})}{\|p_k\|(\|q^*(r_k) - x_k\| + \|s^*(r_k) - x_k\|)} \leq \frac{(r_k, d_k^{AW} + d_k^{FW})}{\|p_k\|(\|d_k^{AW}\| + \|d_k^{FW}\|)} \leq 2\frac{(r_k, d_k^{FW})}{\|p_k\|(\|d_k^{FW}\| + \|d_k^{AW}\|)} \leq \\
\leq&2\frac{(r_k, d_k^{FW})}{\|p_k\|\|d_k^{FW}\|} \leq 2\frac{1+\beta}{1-\beta}\frac{(\tilde{d}_k^{FW}, r_k)}{\|p_k\|\|\tilde{d}_k^{FW}\|}
\end{aligned} \tag{2.5.72}$$

Therefore since the algorithm in Table 3 sets $\tilde{d}_k = \tilde{d}_k^{FW}$ if $\tilde{d}_k^{AW}$ is not computed

$$\frac{1-\beta}{2(1+\beta)}\|p_k\|\mathrm{NW}^f(\Omega) \leq \frac{(\tilde{d}_k^{FW}, r_k)}{\|\tilde{d}_k^{FW}\|} = \frac{(\tilde{d}_k, r_k)}{\|\tilde{d}_k\|} \tag{2.5.73}$$

Summarizing, we have that in case 1 the approximated algorithm has indeed the same descent property of the exact one described by equation (2.5.30) with $\frac{1-\beta}{2+2\beta}\mathrm{NW}^f(\Omega)$ instead of $\mathrm{NW}^f(\Omega)$.

Case 2: $\alpha_k = \alpha_{\max} = 1$, $\tilde{d}_k = \tilde{d}_k^{FW}$. On the one hand combining (2.5.32) with (2.5.61) we get

$$\tilde{h}_k \leq (-\nabla f(x_k), d_k) \leq (1+\beta)(-\nabla f(x_k), \tilde{d}_k) \tag{2.5.74}$$

while on the other hand (2.5.34) still holds for the approximated algorithm

$$\tilde{h}_k - \tilde{h}_{k+1} \geq \frac{1}{2}(-\nabla f(x_k), \tilde{d}_k) \tag{2.5.75}$$

Multiplying the second inequality by $2(1+\beta)$ and concatenating

$$(2+2\beta)(\tilde{h}_k - \tilde{h}_{k+1}) \geq \tilde{h}_k \tag{2.5.76}$$

which can be rewritten as

$$\tilde{h}_{k+1} \leq \frac{1+2\beta}{2+2\beta}\tilde{h}_k \tag{2.5.77}$$

Hence the same linear descent property of the exact algorithm holds with $\frac{1+2\beta}{2+2\beta}$ instead of $\frac{1}{2}$.
Case 3: $\alpha_k = \alpha_{\max}$, $\tilde{d}_k = \tilde{d}_k^{AW}$ The analysis does not change in this case since we only need $\tilde{d}_k$ to be a descent direction.
Given these analogies in the analysis of the 3 possible kinds of steps, the rest of the proof is identical to the one of Theorem (2.5.2). □

Chapter 3

Active set complexity

3.1 Introduction

In this chapter we discuss the complexity of the active set problem. We first introduce the problem from a geometric and an algebraic point of view, in terms of exposed faces and non zero Lagrangian multipliers respectively. We then show that for polytopes these two approaches yield equivalent definitions. In the rest of the chapter we makes several examples of methods with the finite time active set identification property. While in this chapter our analysis is done in the geometric framework, in chapter 4 we will analyze the AFW active set complexity using the algebraic framework.

3.1.1 Exposed faces

In the rest of this chapter Ω will be a convex and closed subset of $\mathbb{R}^n$, $f : \Omega \to \mathbb{R}$ differentiable with X^* the set of local minima for f.
We first recall the definition of exposed face:

Definition 3.1.1. If Ω is a closed convex set and c a linear function the face of Ω exposed by c is the the set

$$E_\Omega(c) = \text{argmax}\{cx \mid x \in \Omega\} \tag{3.1.1}$$

It follows immediately from the definition that $x \in E_\Omega(c)$ if and only if $c \in N_\Omega(x)$ (see for instance [11] for a proof). Since the first order optimality conditions can be expressed as $-\nabla f(x^*) \in N_\Omega(x^*)$, they can also equivalently be written as $x^* \in E(-\nabla f(x^*))$. We can now define the support of a subset of solutions:

Definition 3.1.2. We will say that a face $\mathcal{F}$ of Ω is the geometric support of a subset A of X^* with respect to f and write $\mathcal{F} = \mathcal{A}_f(A)$ if

$$\mathcal{F} = E_\Omega(-\nabla f(x)) \tag{3.1.2}$$

for every $x \in A$.

When it is clear from the context what function we are considering we will simply say that $\mathcal{F}$ is the geometric support of A.

Notice that $A \subset \mathcal{F}$ because by first order optimality conditions $x \in E_\Omega(-\nabla f(x))$ for every $x \in X^*$. This notion of support allows us to formally define what it means to solve the active set problem for sequences converging to the set of minimizers.

Definition 3.1.3. Let $\{x_k\}_{k\in\mathbb{N}} \subset \Omega$ be a sequence converging to a subset A of X^* with geometric support $\mathcal{F}$, that is to say

$$\mathrm{dist}(x_k, A) \to 0 \tag{3.1.3}$$

with $\mathcal{A}_f(A) = \mathcal{F}$. We say that $\{x_k\}_{k\in\mathbb{N}}$ solves the active set problem in M steps if $x_k \in \mathcal{F}$ for every $k \geq M$.

We now introduce the definition of polyhedral face, which comes from a generalization of a property concerning the faces of polyhedral sets.

Definition 3.1.4. A face $\mathcal{F}$ of Ω is said to be polyhedral if for any $x \in \mathrm{ri}(\mathcal{F})$:

$$\mathrm{aff}(\mathcal{F}) = \{x\} + \mathrm{lin}(T_\Omega(x)) \tag{3.1.4}$$

As a notable example, all the faces of a polyhedral set are polyhedral. We now need to introduce two properties of polyhedral faces that define the structure of their normal cones. These properties of normal cones will be relevant when studying the projection on Ω, because $\pi_\Omega(x) = y \Leftrightarrow x \in \{y\} + N_\Omega(y)$.

Proposition 3.1.5. *Let $\mathcal{F}$ be a polyhedral face of Ω. Then for every $x, y \in \mathrm{ri}(\mathcal{F})$ we have $N_\Omega(x) = N_\Omega(y)$.*

Proof. See [12]. □

Thanks to this first property one can define $N(\mathcal{F}) = N_\Omega(x)$ for some $x \in \mathrm{ri}(N(\mathcal{F}))$, and the definition does not depend on x. We can now state the second property:

Proposition 3.1.6. *Let $\mathcal{F}$ be a polyhedral face of Ω. Then*

1. *$d \in \mathrm{ri}(N(\mathcal{F}))$ if and only if $E_\Omega(d) = \mathcal{F}$.*
2. *For every $x \in \mathcal{F}$ the cone $N(\mathcal{F})$ is a face of $N_\Omega(x)$.*

Proof. 1. See [12].
2. If $\mathcal{F}$ is a singleton the statement is trivial because $N(x) = N(\mathcal{F})$ for the only point $x \in \mathcal{F}$. Otherwise consider $y \neq x$ such that $y \in \mathrm{ri}(\mathcal{F})$. Let $e = y - x$. We claim that $N(\mathcal{F}) = E_{N(x)}(e)$. Since $(e, d) = (y - x, d) \leq 0$ for every $d \in N_\Omega(x)$ we have

$$\max_{d\in N_\Omega(x)}(e, d) = 0 \tag{3.1.5}$$

so that $d \in E_{N(x)}(e)$ if and only if $(d, e) = 0$. We now prove the two inclusions.
$\subseteq$: for every $d \in N(\mathcal{F}) \subset N_\Omega(x)$

$$(d, x) = (d, y) = \max_{z\in\Omega}(d, z) \tag{3.1.6}$$

so that in particular $(d, e) = (d, y - x) = 0$ which means $d \in E_{N(x)}(e)$.
$\supseteq$: if $(d, e) = 0$ then again (3.1.6) and therefore $d \in N_\Omega(y) = N(\mathcal{F})$. □

3.1.2 Optimality conditions

We now give an algebraic definition of active set complexity. This complexity will be defined for constrained problems as the number of iterations that it takes to identify a certain subset of the constraints which are satisfied on a possibly local minimum. This subset corresponds to the indexes of possibly non zero Lagrangian multipliers in some KTT like optimality conditions. There are of course different kind of hypotheses that imply the existence of Lagrangian multipliers, so that it is now convenient to pick one type of conditions which encompasses all the problems we will deal with in this chapter.
In the rest of this chapter for any constrained problem and any feasible point x^* we call $\mathcal{A}(x^*)$ the subset of indexes of the inequality constraints which are active (satisfied with equality) in x^*.
We now recall stationarity conditions for systems of inequalities (see for instance [25] for a proof).

Proposition 3.1.7. *Given the problem*

$$\min\{f(x) \mid g_i(x) \leq 0 \ \forall\ 1 \leq i \leq n, \quad h_j(x) = 0 \ \forall\ 1 \leq j \leq m\} \tag{3.1.7}$$

with $g_i(x)$ convex and differentiable, $h_i(x)$ affine assume that x^* is a local constrained minimum and that f, g_i are differentiable in x^*. Assume also and there exists $\bar{x}$ such that $h_i(\bar{x}) = 0$, $g_i(\bar{x}) < 0$ (this condition is known as SMFCQ). Then there exists $\lambda \in \mathbb{R}^n$, $\mu \in \mathbb{R}^m$ such that

$$\begin{aligned} \nabla f(x) + \lambda^T \nabla g(x^*) + \mu^T \nabla h(x^*) &= 0 \\ \lambda^* &\geq 0 \\ (g(x^*), \lambda) &= 0 \end{aligned} \tag{3.1.8}$$

where $g(x) = (g_1(x), ..., g_n(x))^T$ and $h(x) = (h_1(x), ..., h_n(x))^T$. When f is convex, the converse is also true.

We can finally define the set of proper active constraints $\mathcal{A}^+(x^*)$.

Definition 3.1.8. Under the assumptions of Proposition 3.1.7, we say that an index $i \in \{1, ..., n\}$ is in $\mathcal{A}^+(x^*)$ if there exists Lagrangian multipliers (λ, μ) satisfying (3.1.8) such that $\lambda_i > 0$.

By convexity and positive linearity of the optimality conditions, it is easy to see that there exists λ satisfying (3.1.8) such that $\lambda_i > 0$ for every $i \in \mathcal{A}^+(x^*)$.
Finally, we define the algebraic support of a subset of solutions:

Definition 3.1.9. Let $\mathcal{A}^+$ be a subset of $\{1, ..., n\}$ and A be a subset of X^*. We say that the surface $\mathcal{F} = \{x \in \Omega \mid g_i(x) = 0 \ \forall i \in \mathcal{A}^+\}$ is the algebraic support of A and write $\mathcal{F} = \mathcal{A}^a_f(A)$ if $\mathcal{A}^+(x) = \mathcal{A}^+$ for every $x \in A$.

In the rest of this section we cite a few results concerning the connection between active sets of constraints and normal or tangent cones.
The following theorem which relates the normal cone to the active constraints is a particular case of [24], Theorem 3:

Theorem 3.1.10. *Under the hypotheses of Proposition 3.1.7 if Ω is the feasible set then*

$$T_\Omega(x^*) = \{d \in \mathbb{R}^n \mid (\nabla g_i(x^*), d) \leq 0 \ \forall\ i \in \mathcal{A}(x^*),\ (\nabla h_i(x^*), d) = 0 \ \forall\ 1 \leq i \leq m\} \tag{3.1.9}$$

Then this description of the normal cone immediately follows by linear duality:

Corollary 3.1.11. *Under the hypotheses of Proposition 3.1.7 if Ω is the feasible set then*

$$N_\Omega(x^*) = \{ \sum_{i \in \mathcal{A}(x^*)} \lambda_i \nabla g_i(x^*) + \sum_{i=1}^{m} u_i \nabla h_i(x^*) \mid \lambda_i \geq 0 \ \forall\ i \in \mathcal{A}(x^*)\} \tag{3.1.10}$$

As for $\mathcal{A}^+(x^*)$, it turns out to be a set of generators together with $\{\nabla h_j(x^*)\}_{j \in \{1,\ldots,m\}}$ of the smallest face of $N_\Omega(x^*)$ containing $-\nabla f(x^*)$.

Proposition 3.1.12. *Under the hypotheses of Proposition 3.1.7 the set*

$$\mathcal{F} = \mathrm{cone}(\{\nabla g_i(x^*)\}_{i \in \mathcal{A}^+(x^*)}) + \mathrm{span}(\{\nabla r_i\}_{1 \leq i \leq m})$$

is a face of $N_\Omega(x^*)$ and $-\nabla f(x^*) \in \mathrm{ri}(\mathcal{F})$.

Proof. Follows from Proposition 5.1.5 of the appendix, considering as set of generators

$$G = \{\nabla g_i(x^*)\}_{i \in \mathcal{A}^+(x^*)} \cup \{\nabla r_i\}_{1 \leq i \leq m} \cup \{-\nabla r_i\}_{1 \leq i \leq m}$$

□

These results allow us to study the relation between algebraic and geometric support.

3.1.3 Equivalence of definitions for linear constraints

In general for any $x \in X^*$ the geometric support is a subset of the algebraic support, and the inclusion can be strict as can be seen for instance in balls. Indeed given the euclidean unit ball described by the constraint $\|x\|^2 \leq 1$ and a differentiable function $f : B(0,1) \to \mathbb{R}$ having a non singular minimum in $p \in \partial B(0,1)$ it is easy to check that $\mathcal{A}_f(\{p\}) = \{p\}$ while $\mathcal{A}_f^a(\{p\}) = \partial B(0,1)$. We now prove the inclusion.

Proposition 3.1.13. *Under the hypotheses of Proposition 3.1.7 for every $\bar{x} \in X^*$*

$$E_\Omega(-\nabla f(\bar{x})) \subseteq \mathcal{A}_f^a(\{\bar{x}\})$$

Proof. For every $x \in \Omega$, $i \in \mathcal{A}(\bar{x}) \supseteq \mathcal{A}^+(\bar{x})$ we have

$$(\nabla g_i(\bar{x}), (x - \bar{x})) \leq 0 \tag{3.1.11}$$

because $g_i(\bar{x}) = 0$, $g_i(x) \leq 0$ and $g_i(x)$ is convex. By hypothesis we can apply Proposition 3.1.7 to obtain

$$-\nabla f(\bar{x}) = \sum_{i \in \mathcal{A}^+(\bar{x})} \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^{m} u_j \nabla h_j(\bar{x})$$

where we can choose $\lambda_{\mathcal{A}^+(\bar{x})}$ such that $\lambda_i > 0$ for every $i \in \mathcal{A}^+(\bar{x})$. Let $x \in E_\Omega(-\nabla f(\bar{x}))$, so that $(x - \bar{x}, \nabla f(\bar{x})) = 0$. Then

$$\begin{aligned} 0 = (-\nabla f(\bar{x}), x - \bar{x}) &= \sum_{i \in \mathcal{A}^+(\bar{x})} \lambda_i (\nabla g_i(\bar{x}), x - \bar{x}) + \sum_{j=1}^{m} u_j (\nabla h_j(\bar{x}), x - \bar{x}) = \\ &= \sum_{i \in \mathcal{A}^+(\bar{x})} \lambda_i (\nabla g_i(\bar{x}), x - \bar{x}) \end{aligned} \tag{3.1.12}$$

Applying (3.1.11) to this equation we obtain that the last sum is equal to 0 if and only if $(\nabla g_i(\bar{x}), x - \bar{x}) = 0$ for every $i \in \mathcal{A}^+(\bar{x})$. We then have by convexity $g_i(x) \geq 0$ for every $i \in \mathcal{A}^+(\bar{x})$ forcing $g_i(x) = 0$ for every $i \in \mathcal{A}^+(\bar{x})$ and in particular $x \in \mathcal{A}^a_\Omega(\{\bar{x}\})$. □

When the $\{g_i\}_{i \in \mathcal{A}^+(\bar{x})}$ are linear, positive multipliers are related to the face *exposed by* the negative gradient $-\nabla f(x)$. We begin the proof with the following lemma concerning a feasible set given by the intersection of two closed convex sets.

Lemma 3.1.14. *Let $\Omega = P \cap U$ where P, U are closed convex sets. Let $x \in P$ and $d \in N_P(x)$. Then if $x \in \Omega$*

$$E_\Omega(d) = E_P(d) \cap U \tag{3.1.13}$$

Proof. The hypothesis $d \in N_P(x)$ can be equivalently rewritten as $x \in E_P(d)$. Then

$$(x, d) = \max\{y \in P \mid (y, d)\} \tag{3.1.14}$$

and since $\Omega = P \cap U \subset P$, $x \in \Omega$

$$(x, d) = \max\{y \in \Omega \mid (y, d)\} \tag{3.1.15}$$

so that

$$\begin{aligned} E_\Omega(d) = \operatorname{argmax}\{(y, d) \mid y \in \Omega\} = \{y \in \Omega \mid (y, d) = (x, d)\} = \\ = \{y \in P \mid (y, d) = (x, d),\ y \in U\} = E_P(d) \cap U \end{aligned} \tag{3.1.16}$$

□

Proposition 3.1.15. *For a constrained problem like the one in 3.1.7, let $\bar{x} \in X^*$ and assume that g_i is affine for every $i \in \mathcal{A}^+(\bar{x})$. Than the geometric support of $\{\bar{x}\}$ coincides with the algebraic support of $\{\bar{x}\}$:*

$$E_\Omega(-\nabla f(\bar{x})) = \{x \in \Omega \mid g_i(x) = 0\ \forall\ i \in \mathcal{A}^+(\bar{x})\} \tag{3.1.17}$$

Proof. Let P be the polyhedral set defined by

$$P = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\ \forall\ i \in \mathcal{A}^+(x),\ h_j(x) = 0\ \forall\ 1 \leq j \leq m\}$$

Then by Proposition 3.1.12 we have $-\nabla f(\bar{x}) \in \operatorname{ri}(N_P(\bar{x}))$. Since $\bar{x}$ satisfies all constraints with equality the minimal face of P containing x is

$$\mathcal{F}_P(\bar{x}) = \{x \in \mathbb{R}^n \mid g_i(x) = 0\ \forall\ i \in \mathcal{A}^+(x), r_j(x) = 0\ \forall\ 1 \leq j \leq m\}$$

which is an affine subspace. In particular $\mathcal{F}_P(\bar{x}) = \mathrm{ri}(\mathcal{F}_P(\bar{x}))$ so that $\bar{x} \in \mathrm{ri}(\mathcal{F}_P(\bar{x}))$. By Lemma 3.1.6, this together with $-\nabla f(\bar{x}) \in \mathrm{ri}(N_P(\bar{x}))$ implies $E_P(-\nabla f(\bar{x})) = \mathcal{F}_P(\bar{x})$. Let now $U = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0 \ \forall\ i \in \{1, ..., n\} \setminus \mathcal{A}^+(\bar{x})\}$. Then

$$\begin{aligned}\mathcal{A}_f^a(\bar{x}) &= \{x \in \Omega \mid g_i(x) = 0 \ \forall\ i \in \mathcal{A}^+(\bar{x})\} = \\ &= U \cap \{x \in P \mid g_i(x) = 0 \ \forall i \in \mathcal{A}^+(\bar{x}), \nabla h_j(\bar{x}) = 0 \ \forall\ 1 \leq j \leq m\} = \\ &= U \cap \mathcal{F}_P(\bar{x}) = U \cap E_P(-\nabla f(\bar{x}))\end{aligned} \tag{3.1.18}$$

where by Lemma 3.1.14 since $\Omega = U \cap P$

$$U \cap E_P(-\nabla f(\bar{x})) = E_\Omega(-\nabla f(\bar{x}))$$

□

We have the following corollary for polyhedra, that is when $\Omega = P$ in the notation of Proposition 3.1.15.

Corollary 3.1.16. *If Ω is a polyhedron, the support and the algebraic support of any subset of X^* coincide when one of the two exists.*

Proof. Just apply the previous proposition to any point a of a set $A \subset X^*$ with constant support. □

3.2 A new proof for PG finite time active set identification with convergence estimates

In this section we present a new proof of finite time active set identification for the projected gradient method on polytopes and give explicit convergence estimates under suitable hypotheses on f. P will be a polyhedron and given $x \in P$ the set $\mathcal{F}(x)$ will be the minimal face of P containing x. This proof differs from the one [12] because we will not use normal cones to faces and the corresponding partition lemma. We will instead do a local analysis which will highlight how the active set radius depend on the position of the optimum and on the gradient of the objective function on the optimum.
As it was done in the original proof (see [11], [12]) we start by recalling a few relevant properties of convex sets. We begin by characterizing the minimal face containing a certain point x.

Lemma 3.2.1. *For every $x \in P$ the minimal face $\mathcal{F}(x)$ of P containing x is uniquely defined by $x \in \mathrm{ri}(\mathcal{F}(x))$.*

Proof. Let

$$\mathcal{F}(x) = \cap_{\substack{G \text{ is a face of } P \\ x \in G}} G$$

be the minimal face of P containing x. Assume by contradiction $x \notin \mathrm{ri}(G)$. Then by the separation theorems there would exists an hyperplane $(c, y) = (c, x)$ separating x from $\mathrm{ri}(\mathcal{F})$. But the intersection between this hyperplane and $\mathcal{F}$ would be a smaller face of P containing x, absurd. □

In the above proof we used implicitly that all the faces of $\mathcal{F}(x)$ are still faces of P, which is well known (see for instance [17], proposition 3.25). A corollary of Lemma (3.2.1) is the partition property (see for instance [41] for a proof)

$$P = \bigcup_{\mathcal{F} \text{ is a face of P}} \mathrm{ri}(\mathcal{F}) \tag{3.2.1}$$

We now need a technical lemma in which we relate the width of P with respect to tangent cone in a fixed point with the distance of this point from a certain set of faces. In practice, this lemma will allow us to consider $T_P(x)$ instead of P in a small enough neighborhood of x.

Lemma 3.2.2. *For $x \in P$ let $r = l^b_{P,x}(0,1)$ defined in chapter 1. Consider the set $\mathcal{G}$ of faces of P such that $\mathcal{F}(x)$ is not a subset of G. If*

$$D = \bigcup_{H \in \mathcal{G}} H$$

then $\mathrm{dist}(x, D) = r > 0$

Proof. Any face H in $\mathcal{G}$ can not contain x otherwise we would have $\mathcal{F}(x) \subset H$ by the minimality of $\mathcal{F}(x)$.
It is thus clear that $\mathrm{dist}(x, D) > 0$, since D is a closed set not containing x.
Now just applying the definition we get

$$r = l^b_{P,x}(0,1) = \inf\{l_{P,x}(\hat{c}) \mid c \in T_P(x)\} \tag{3.2.2}$$

We will first show that $r \leq \mathrm{dist}(x, D)$. Let p be a projection of x on the set D, and let $c = (p - x)$ so that $c \in T_P(x)$. Then $l_{P,x}(\hat{c}) \geq \|p - x\|$ because $p \in P$. Moreover, there exists by hypothesis a (proper) face of P containing p but not x, so that $x + \lambda(p - x)$ is not in P for every $\lambda > 1$. To see this, consider $(q, \cdot)$ a linear function exposing the face of P containing p but not x. Then on the one hand

$$(q, p) = \max\{y \in P \mid (q, y)\} \tag{3.2.3}$$

and on the other hand

$$(q, x) < (q, p) \Rightarrow (q, \lambda(p - x)) > (q, p) \text{ for } \lambda > 1 \tag{3.2.4}$$

which proves $x + \lambda(p - x) \notin P$ for $\lambda > 1$.
We can now deduce $l_{P,x}(\hat{c}) = \|p - x\| = \mathrm{dist}(x, D)$ and the $\leq$ is proved by (3.2.2).
It remains to prove $r \geq \mathrm{dist}(x, D)$, or in other words $l_{P,x}(\hat{c}) \geq \mathrm{dist}(x, D)$ for every $c \in T_P(x)$. For a fixed $\hat{c}$ let $\lambda_c = l_{P,x}(\hat{c})$ so that $y = x + \lambda_c \hat{c} \in P$, $x + \lambda \hat{c} \notin P$ for every $\lambda > \lambda_c$. If we prove $y \in D$ we are done because then clearly

$$l_{P,x}(\hat{c}) = \lambda_c = \|y - x\| \geq \mathrm{dist}(x, D) \tag{3.2.5}$$

Assume by contradiction that $y \notin D$, or equivalently $\mathcal{F}(x) \subset \mathcal{F}(y)$. Since $y \in \mathrm{ri}(\mathcal{F}(y))$ by definition and $x \in \mathcal{F}(y)$ we would then have $x + \lambda(y - x) \in \mathcal{F}(y) \subset P$ for some $\lambda > 1$, contradiction with the maximality of λ_c. Then $y \in D$ and we are done by (3.2.5). □

We now apply the lemma we just proved to show the local coincidence of P and $T_P(x)$. Furthermore, we show with quantitative estimates the lower semicontinuity of $\mathcal{F}(x)$ on P as a set valued function.

Lemma 3.2.3. *For every $x \in P$ let $r = l^b_{P,x}(0,1)$ as in Lemma 3.2.2. Then:*

1. $P \cap B(x,r) - \{x\} = T_P(x) \cap B(0,r)$
2. *For every $y \in P \cap B(x,r)$ we have $\mathcal{F}(y) \supset \mathcal{F}(x)$.*

Proof. 1. Since $P - \{x\} \subseteq T_P(x)$ we have $P \cap B(x,r) - \{x\} \subseteq T_P(x) \cap B(x,r)$. On the other hand for every $c \in T_P(x) \setminus \{0\}$ there exists $y \in P$ such that $\hat{c} = \frac{y-x}{\|y-x\|}$. Reasoning as in Lemma 3.2.2 we get $x + \lambda \hat{c} \in P$ for every $0 \le \lambda \le r$. If $c \in T_P(x) \cap B(0,r) \setminus \{0\}$ we have $\|c\| < r$ and $c = \|c\|\hat{c}$, so that in particular $c \in P \cap B(x,r) - \{x\}$.
2. This is a corollary of Lemma 3.2.2, since using the notation introduced in the lemma $r = \operatorname{dist}(x, D)$ with $y \notin D$ if and only if $\mathcal{F}(y) \supset \mathcal{F}(x)$ for $y \in P$. $\square$

Having proved these facts we can now describe how the projection on a cone behaves in a neighborhood of a point in the dual cone.

Proposition 3.2.4. *Let C be a polyhedral convex cone and let $e \in C^d$. Let $\mathcal{F}(e)$ be the minimal face of C^d containing e. Let*

$$r = l^b_{C^d,e}(0,1)$$

Then for every $x \in B(e,r)$ the projection $\pi_C(x)$ of x on C is on $E_C(e)$.

The proof relies on the Moreau Yosida decomposition and on the duality Lemma 5.1.6. In this proof and in the rest of this section we use the notation $\pi_A(x)$ for the projection of x on a closed convex set A.

Proof. Let $x \in B(e,r)$. By the Moreau Yosida decomposition

$$x = \pi_C(x) + \pi_{C^d}(x) \text{ with } \pi_C(x) \perp \pi_{C^d}(x) \tag{3.2.6}$$

Since the projection is 1 - Lipschitz, we have $\pi_{C^d}(x) \in B(e,r)$ and by Lemma 3.2.3 we have $\mathcal{F}(\pi_{C^d}(x)) \supset \mathcal{F}(e)$.
Since $\pi_C(x) \perp \pi_{C^d}(x)$ we have, using the notation of Lemma 5.1.6:

$$\pi_C(x) \in (\mathcal{F}(\pi_{C^d}(x))^* \subset (\mathcal{F}(e))^* \subset E_C(e) \tag{3.2.7}$$

where we recall that $(\cdot)^*$ reverse inclusions and in the last inclusion we are using $E_C(e) = C \cap e^\perp$. $\square$

We can finally describe quantitatively how close a point must be to a certain vector in the normal cone to identify the same face in a polyhedral set:

Lemma 3.2.5. *Let P be a polyhedral set, let $x \in P$ and let $e \in N_P(x)$. Let*

$$r_e = l^b_{N_P(x),e}(0,1), \;\; r_x = l^b_{P,x}(0,1) \tag{3.2.8}$$

and $r_m = \min(r_e, r_x)$ Then for every $y \in B(e, r_m)$ the projection $\pi_P(y+x)$ of $y+x$ on P is on $E_P(e)$.

Proof. Let $p = \pi_P(y)$, $p' = \pi_P(T_P(x))$. On the one hand by Lemma 5.1.10 we have $\|p'\| = \operatorname{dist}(y, N_P(x)) < r_m < r_x$ so that $p' \in T_P(x) \cap B(0, r_x) = P \cap B(x, r_x) - \{x\}$ where the last equality is justified by Lemma 3.2.3 . Then $p = p' + x$ with $p' \in E_{T_P(x)}(e)$ by Lemma 3.2.4. This implies $(e, x) = (e, p' + x) = (e, p)$, so that also $p \in E_P(e)$. $\square$

Given Lemma 3.2.5 it is straightforward to compute how close the sequence generated by a projected gradient method must be to the optimal point for the active set identification to happen.

Lemma 3.2.6. *Suppose that $x^* \in P$ satisfies first order optimality conditions for a function f with $L-$ Lipschitz gradient so that $-\nabla f(x^*) \in N_P(x^*)$. Let $r_* = l^b_{P,x^*}(0,1)$ and $r_\nabla = l^b_{N(x^*),-\nabla f(x^*)}(0,1)$. For every $x \in P$ and $\alpha > 0$ such that*

$$(1 + \alpha L)\|x - x_*\| < \min(\alpha r_\nabla, r_*) \tag{3.2.9}$$

we have $\pi_P(x - \alpha \nabla f(x)) \in E_P(-\nabla f(x^*))$

In the rest of this proof we use $N(\cdot)$ as a shorthand for $N_P(\cdot)$.

Proof. We have by the Lipschitz condition

$$\begin{aligned}\|(x - \alpha\nabla f(x)) - (x^* - \alpha \nabla f(x^*))\| \leq \|x - x^*\| + \alpha\|\nabla f(x) - \nabla f(x^*)\| \leq \\ \leq (1 + \alpha L)\|x - x^*\| < \min(\alpha r_\nabla, r_*)\end{aligned} \tag{3.2.10}$$

Now we notice that

$$\alpha r_\nabla = \alpha l^b_{N(x^*),-\nabla f(x^*)}(0,1) = l^b_{N(x^*),-\alpha\nabla f(x^*)}(0,1) \tag{3.2.11}$$

as it is immediate to check from the definitions using that $N(x^*)$ is a cone. To conclude now it suffice to apply Lemma 3.2.5 with $x - \alpha\nabla f(x)$ instead of y and $(x^*, -\alpha\nabla f(x^*))$ instead of (x, e). $\square$

We can now apply this lemma to the well known results about projected gradient method convergence to get finite time active set identification results:

Theorem 3.2.7. *Under the assumptions of 3.2.6, suppose additionally that $f(x)$ is convex. If $\{x_k\}_{k\in\mathbb{N}}$ is a sequence generated by the projected gradient method with decreasing step size $\{\alpha_k\}_{k\in\mathbb{N}} \to \bar{\alpha}$ then*

1. *There exists a minimizer $x^* \in P$ of f such that $\{x_k\} \to x^*$*
2. *For r_* and r_∇ defined as in Lemma 3.2.6 we have $x_k \in E_P(-\nabla f(x^*))$ for every $k \geq \bar{k}$, where $\bar{k}$ is the minimum index such that*

$$(1 + \alpha_{\bar{k}} L)\|x_{\bar{k}} - x_*\| < \min(\bar{\alpha} r_\nabla, r_*) \tag{3.2.12}$$

Proof. By [6], proposition 6.1.7 there exists a minimizer x^* for f such that $x_k \to x^*$ and $\|x_k - x^*\| \to 0$ is decreasing in k. So 1. follows immediately and 2. follows by 3.2.6. Indeed for every $k \geq \bar{k}$

$$\begin{aligned}(1 + \alpha_k L)\|x_k - x_*\| \leq (1 + \alpha_{\bar{k}} L)\|x_{\bar{k}} - x_*\| < \min(\bar{\alpha} r_\nabla, r_*) \leq \\ \leq \min(\alpha_k r_\nabla, r_*)\end{aligned} \tag{3.2.13}$$

so that condition (3.2.9) is satisfied. $\square$

In the next theorem we state active set identification bounds for strongly convex objectives.

Theorem 3.2.8. *Under the assumptions of 3.2.6, suppose additionally that $f(x)$ is strongly convex, and let x^* be the unique minimizer for f over P. If $\{x_k\}_{k\in\mathbb{N}}$ is a sequence generated by the projected gradient method with constant step size $\bar{\alpha} \in (0, 2/L)$ then*

1. *Let $q = \max(1 - \bar{\alpha}L, 1 - \bar{\alpha}\mu)$ with μ the strong convexity constant of f. Then*

$$\|x_k - x^*\| \leq q^k \|x_0 - x^*\|$$

2. *For r_* and r_∇ defined as in Lemma 3.2.6 we have $x_k \in E_P(-\nabla f(x^*))$ for every $k \geq \bar{k} + 1$, with $\bar{k}$ defined by*

$$\bar{k} = \lceil \frac{\ln((1 + \bar{\alpha}L)(\|x_0 - x^*\|)) - \ln(\min(\bar{\alpha}r_\nabla, r_*))}{\ln(1/q)} \rceil \tag{3.2.14}$$

Proof. 1. Follows by [6], proposition 6.1.8. It is then straightforward to check that for every $k \geq \bar{k}$ the condition (3.2.9) of Lemma 3.2.6 is satisfied as we have already done in 3.2.7. □

3.2.1 Descent directions

We now state a few key facts related to first order algorithms that will be useful in the analysis of the active set complexity problem. In the rest of this section Ω is a closed compact set and $f : \Omega \to \mathbb{R}$ is a differentiable function with gradient having Lipschitz constant L.
First, we will define the descent directions and the $\lambda_{\max}$ function that are fundamental in the analysis of first order algorithms especially when the line search method is employed to compute the step size.

Definition 3.2.9. Given $x \in \Omega$, $d \in \mathbb{R}^n$ such that $x + d \in \Omega$ we say that d is a descent direction and write $d \in D_\Omega(x)$ if $f(x + d) < f(x + td)$ for every $t \in [0, 1)$.
We define $\lambda_{\max}(\Omega, x, d) = \max\{\lambda \in \mathbb{R}_{\geq 0}\}x + \lambda d \in \Omega\}$.

When the objective function is convex we can determine whether a direction is a descent direction by studying its subdifferential. In the following proposition we use the notation (A, x) with $A \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$ to denote the set $\{(x, a) \mid a \in A\}$.

Proposition 3.2.10. *Let $g : \Omega \to \mathbb{R}$ be a convex function. If $x, x + d \in \Omega$ and*

$$\mathbb{R}_{<0} \cap \partial f(x + d) \cdot d \neq \emptyset$$

then $d \in D_\Omega(x)$.

Proof. By hypothesis there exists $z \in \partial f(x + d)$ such that $(d, z) < 0$. Then since $z \in \partial f(x + d)$

$$f(x + td) \geq f(x + d) + (x + td - (x + d), z) = f(x + d) + (t - 1)(d, z) > f(x + d)$$

for every $t \in [0, 1)$. □

If f has Lipschitz gradient then the usual quadratic upper bound and descent lemma properties are easy to relate to descent directions:

Proposition 3.2.11. *Assume that f has Lipschitz gradient with constant L.*

1. $f(x+d) \leq f(x) + (\nabla f(x), d) + \frac{L}{2}\|d\|^2$
2. $(\nabla f(x+d), d) \geq (\nabla f(x), d) - L\|d\|^2$
3. *If* $(\nabla f(x), d) - L\|d\|^2 \leq 0$ *then d is a descent direction.*

Proof. 1. See [6], proposition 6.1.2.
2. Follows immediately from the Lipschitz property of f

$$(\nabla f(x+d) - \nabla f(x), d) \geq -\|\nabla f(x+d) - \nabla f(x)\|\|d\| \geq -L\|d\|^2 \tag{3.2.15}$$

3. Follows from Lemma 3.2.10 and point 2.

□

3.3 Active set identification for AFW and PFW

In this section we prove finite time active set identification for the AFW and the PFW algorithm on polytopes assuming (continuous) differentiability on the objective and convergence to a subset of minimizers with constant support. In [9] it has already been proved that for the AFW and the PFW methods on the simplex the active set is identified in finite time with a few additional assumptions on the objective function but without the convergence assumptions that we use here.
As for the projected gradient method, we start by analyzing a few relevant properties of closed convex cones. We use these properties to prove that any method employing search directions transversal to the active set and "close" to its tangent cone eventually does maximal steps if it hasn't already identified the active set. This then becomes the key lemma in the proofs concerning the AFW and the PFW methods.
Given a cone C it is well known the dual cone C^d can always be decomposed in an orthogonal sum

$$C^d \cap \mathrm{aff}(C) \bigoplus \mathrm{aff}(C)^\perp = C^d \cap \mathrm{aff}(C) \bigoplus \mathrm{lin}(C^d) \tag{3.3.1}$$

By the definition of dual cone, we have $(d, p) = 0$ for any $p \in C^d$ and $d \in C$. In the next two propositions we compute bounds for (d, p) when d is still in C but p is not in C^d. Of course our bounds will depend on how close p is to C^d and to the orthogonal complement of $\mathrm{aff}(C)$. We use the notation $\mathrm{rbd}(A)$ for the relative boundary of a convex set A: $\mathrm{rbd}(A) = A \setminus \mathrm{ri}(A)$.

Proposition 3.3.1. *Let C be a closed and convex cone, $d \in \mathrm{ri}(C)$ and $\delta = \mathrm{dist}(\mathrm{rbd}(C), d)$. For every $p \in C^d$*

$$(d, p) \leq -\delta\|p - \pi_{\mathrm{lin}(C)}(p)\| \tag{3.3.2}$$

Proof. We now apply the decomposition (3.3.1) to write p as an orthogonal sum. We have

$$p = \pi_{\text{lin}(C^d)}(p) + (p - \pi_{\text{lin}(C^d)}(p)) = \pi_{\text{lin}(C^d)}(p) + p_C$$

with $\pi_{\text{lin}(C^d)}(p) \in \text{lin}(C^d) = \text{aff}(C)^\perp$ and $p_C \in \text{aff}(C)$ so that $(d, \pi_{\text{lin}(C^d)}(p)) = 0$. Thus $(p, d) = (d, p_C)$, and by the definition of δ we have $d - \delta\hat{p}_C \in C$, so that $(d, p_C) + \delta\|p\|_C = (d + \delta\hat{p}_C, p_C) \le 0$, and the thesis follows. □

Proposition 3.3.2. *Let C be a closed and convex cone, $d \in \text{ri}(C)$ and $\delta = \text{dist}(\text{rbd}(C), d)$. For every $p \in \partial B(0,1)$, if $\gamma = \text{dist}(p, C^d)$, $\alpha = \|\pi_{\text{lin}(C^d)}(p)\|$ then*

$$(d, p) \le -\delta\sqrt{1 - \gamma^2 - \alpha^2} + \gamma\|d\| \tag{3.3.3}$$

Proof. Since $\mathbb{R}^n = \text{lin}(C^d) \bigoplus \text{aff}(C)$ with $\text{lin}(C^d) \perp \text{aff}(C)$ we can write $p = p_l + p_a$ in a unique way with $p_l \in \text{lin}(C^d)$ and $p_a \in \text{aff}(C)$ so that $p_a \perp p_l$ and

$$1 = \|p\| = \|p_a\|^2 + \|p_l\|^2 = \|p_a\|^2 + \alpha^2 \Rightarrow \|p_a\| = \sqrt{1 - \alpha^2} \tag{3.3.4}$$

Let $\pi_{C^d}(p) = q = q_a + q_l$ with the summands on the right hand side defined analogously to p_a and p_l. Since

$$\|q - p\|^2 = \|q_a - p_a\|^2 + \|q_l - p_l\|^2$$

with the right hand side the minimum for $q \in C^d$ necessarily $q_l = p_l$, so that

$$\|q_a - p_a\| = \text{dist}(p, C^d) = \gamma$$

Since q is a projection on a convex cone with $\lambda q \in C^d$ for every $\lambda \ge 0$ we have

$$p_a - q_a = p - q \perp q \tag{3.3.5}$$

Since $q_a - p_a \in \text{aff}(C)$ we have $q_a - p_a \perp q_l$ and then (3.3.5) implies $q_a - p_a \perp q_a$. Hence

$$1 - \alpha^2 = \|p_a\|^2 = \|q_a\|^2 + \|q_a - p_a\|^2 = \|q_a\|^2 + \gamma^2 \tag{3.3.6}$$

so that $\|q_a\| = \sqrt{1 - \alpha^2 - \gamma^2}$. Applying Proposition 3.3.1 to q we get

$$(d, q) \le -\delta\sqrt{1 - \gamma^2 - \alpha^2} \tag{3.3.7}$$

and since $\|p - q\| = \gamma$

$$(d, p) \le (d, q) + \gamma\|d\| \le -\delta\sqrt{1 - \gamma^2 - \alpha^2} + \gamma\|d\| \tag{3.3.8}$$

□

We now define an identifying property for search directions which generalizes a property of the AFW and PFW search directions.

Definition 3.3.3. Let Ω be a compact convex set and $\mathcal{F}$ be a polyhedral face of Ω. A sequence of directions in $\mathbb{R}^n \setminus \{0\}$ is said to be $\mathcal{F}-$ identifying if for k large enough:

1. $\text{dist}(\hat{d}_k, -T(\mathcal{F})) \to 0$

2. $\|\pi_{\text{aff}(\mathcal{F})}(\hat{d}_k)\| < h$ for some fixed $h < 1$.

We can now prove the lemma which guarantees maximal steps for the AFW and the PFW methods for iterations which have not yet identified the active set.

Lemma 3.3.4. *Let Ω be a compact convex set and $\mathcal{F}$ be a polyhedral face of Ω. Let $f : \Omega \to \mathbb{R}$ be a convex function with continuous differential and let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence in Ω generated doing linesearch along the directions $\{d_k\}_{k\in\mathbb{N}}$. In other words*

$$x_{k+1} \in \operatorname{argmin}\{f(x) \mid x = x_k + \lambda d_k,\ \lambda \geq 0\} \tag{3.3.9}$$

Assume that $\text{dist}(x_k, \mathcal{A}) \to 0$ *for some $\mathcal{A}$ with geometric support $\mathcal{F}$. Then for k large enough $x_{k+1} = x_k + \lambda_{\max}(\Omega, x_k, d_k)$.*

Proof. It suffices to show that $x_{k+1} - x_k \in D_\Omega(x_k)$ for k large enough, or by Lemma 3.2.10 that $(\nabla f(x_{k+1}), x_{k+1} - x_k) < 0$ which is true iff $(\nabla f(x_{k+1}), \hat{d}_k) < 0$. Let a_k be the projection of x_k on $\mathcal{A}$ so that if $\beta_k = \|\nabla f(a_k) - \nabla f(x_k)\|$ then by uniform continuity $\beta_k \to 0$. First, we have inequality

$$(\nabla f(x_{k+1}), \hat{d}_k) \leq (\nabla f(a_k), \hat{d}_k) + \|\nabla f(a_k) - \nabla f(x_k)\| = (\nabla f(a_k), \hat{d}_k) + \beta_k \tag{3.3.10}$$

with $-\nabla f(a_k) \in N(\mathcal{F})$.
Let $\delta_k = \text{dist}(\text{rbd}(N(\mathcal{F})), -\nabla f(a_k))$ so that $\delta_k \geq \delta$ for every k with

$$\delta = \min_{x\in\mathcal{A}} \text{dist}(\text{rbd}(N(\mathcal{F})), -\nabla f(x)) > 0 \tag{3.3.11}$$

because $-\nabla f(x)$ is continuous and $\text{dist}(\text{rbd}(N(\mathcal{F})), -\nabla f(x)) > 0$ for every $x \in \mathcal{A}$ since $\mathcal{F}$ is the geometric support of $\mathcal{A}$.
Let $\alpha_k = \|\pi_{\text{aff}(\mathcal{F})-\{a_k\}}(-\hat{d}_k)\| = \|\pi_{\text{lin}(T(\mathcal{F}))}(-\hat{d}_k)\|$ and $\gamma_k = \text{dist}(T(\mathcal{F}), -\hat{d}_k)$. By hypothesis $\gamma_k \to 0$ and $\alpha_k < h < 1$ for some fixed h. We now apply Lemma 3.3.2 with $d = -\nabla f(a_k)$, $C = T(\mathcal{F})$, $p = -\hat{d}_k$ and obtain

$$(\nabla f(a_k), \hat{d}_k) = (-\nabla f(a_k), -\hat{d}_k) \leq -\delta_k\sqrt{1 - \gamma_k^2 - \alpha_k^2} + \gamma_k\|\nabla f(a_k)\| \tag{3.3.12}$$

Plugging this inequality into (3.3.10) we get

$$\begin{aligned}(\nabla f(x_{k+1}), \hat{d}_k) \leq &- \delta_k\sqrt{1 - \gamma_k^2 + \beta_k^2 - \alpha_k^2} + \gamma_k\|\nabla f(a_k)\| + \beta_k \leq \\ \leq &- \delta\sqrt{1 - \gamma_k^2 - \alpha_k^2} + \gamma_k M + \beta_k\end{aligned} \tag{3.3.13}$$

with $M = \max_{x\in\mathcal{A}} \|\nabla f(x)\|$. To conclude, we have

$$\lim_{k\to\infty} -\delta\sqrt{1 - \gamma_k^2 - \alpha_k^2} + \gamma_k M + \beta_k = -\delta\sqrt{1 - \alpha_k^2} \tag{3.3.14}$$

so that for k large enough $(\nabla f(x_{k+1}), \hat{d}_k) < 0$. □

We recall that given a convex and closed set P, and $x \in P$ we define $\mathcal{F}(x)$ as the minimal face of P containing x. In the next proposition we prove active set finite time identification for an abstract version of the AFW. We do not prove an analogous proposition for the PFW, which does not appear to have a simple geometric generalization.

Proposition 3.3.5. *Let Ω be a compact convex set, $f : \Omega \to \mathbb{R}$ a C^1 function, $\mathcal{A} \subset \Omega$ with geometric support $\mathcal{F}$. Let $\{d_k\}_{k\in\mathbb{N}}$ be a sequence of $\mathcal{F}-$ identifying directions, let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence in Ω and let $M > 0$ be with the following properties:*

1. $\mathrm{dist}(x_k, \mathcal{A}) \to 0$.
2. *For $k \geq M$, either $x_{k+1} \in \mathcal{F}$ or x_{k+1} is the result of a linesearch along d_k from x_k.*
3. *For $k \geq M$, either $x_{k+1} \in \mathcal{F}$ or $x_{k+1} \in \mathcal{F}(x_k)$.*

Then for k large enough $x_k \in \mathcal{F}$.

Proof. By Lemma 3.3.4 applied to property 2. there exists $N \geq M$ such that for $k \geq N$ either $x_k \in \mathcal{F}$ or $x_{k+1} = x_k + \lambda_{\max}(\Omega, x_k, d_k)d_k$.
We first show that if $x_{\bar{k}} \in \mathcal{F}$ for some $\bar{k} \geq N$ then $x_k \in \mathcal{F}$ for every $k \geq \bar{k}$. It suffices to show that $x_{\bar{k}+1} \in \mathcal{F}$ and then the claim follows by induction. But $x_{\bar{k}+1} \in \mathcal{F}$ is immediate because by property 3 x_{k+1} is either in $\mathcal{F}$ or in $\mathcal{F}(x_{\bar{k}}) \subset \mathcal{F}$.
Let $k \geq N$ such that $x_k \notin \mathcal{F}$. Then $x_{k+1} = x_k + \lambda_{\max}(\Omega, x_k, d_k)d_k$ together with $x_{k+1} \in \mathcal{F}(x_k)$ imply that $x_{k+1} \in \mathrm{rbd}(\mathcal{F}(x_k))$ so that $\dim(\mathcal{F}(x_{k+1})) < \dim(\mathcal{F}(x_k))$, which can of course happen finitely many times since the minimal dimension is 0. This means that eventually there must exists $\bar{k}$ such that $x_{\bar{k}} \in \mathcal{F}$, and the theorem follows from the first claim. $\square$

We start to set up the main theorem by introducing an additional assumption which however is not restrictive with respect to the general case. Let $\mathcal{P}$ be a finite set of points in $\mathbb{R}^m$, with $n = |\mathcal{P}|$. We assume $\mathcal{P} = \{e_i\}_{1\leq i\leq n}$ so that $\mathrm{conv}(\mathcal{P}) = \Delta_{n-1} \subset \mathbb{R}^n$. This is not restrictive because the key elements of our theorem are invariant by affine transformation. It is now necessary to introduce some notation to make a more explicit statement and prove it.
Let A be the matrix whose columns A^i are the elements of $\mathcal{P}$, so that $\{Ae_i\}_{1\leq i\leq n} = \mathcal{P}$.
Let $f_P : \mathrm{conv}(\mathcal{P}) \to \mathbb{R}$ with continuous differential and $f : \Delta_{n-1} \to \mathbb{R}$ defined by $f(x) = f_P(Ax)$.
First, if $\{x_k^P\}_{k\in\mathbb{N}}$ in $\mathrm{conv}(\mathcal{P})$ is a sequence generated by the AFW or the PFW with respectively, it is well known (see for instance [28]) that there exists a sequence $\{x_k\}_{k\in\mathbb{N}}$ in Δ_{n-1} generated by the corresponding method applied to f and such that $Ax_k = x_k^P$ for every $x \in \mathbb{N}$.
As for the set of minimizers, it is clear that if X_P^* is the set of minimizers for f_P then $X^* = A^{-1}(X_P^*)$ is the set of minimizers for f on Δ_{n-1}.
If $\mathcal{F}_P$ is a face of $\mathrm{conv}(P)$ then there exists a subset $F_\mathcal{P}$ of $\mathcal{P}$ such that

$$\mathcal{F}_P = \mathrm{conv}(F_\mathcal{P}) = \mathrm{conv}(\{A^i \mid i \in A(\mathcal{F}_P)\}) \tag{3.3.15}$$

for some $A(\mathcal{F}_P) \subseteq \{1, ..., n\}$. We claim that if $\mathcal{F} = A^{-1}(\mathcal{F}_P)$ then

$$\mathcal{F} = \mathrm{conv}(\{e_i \mid i \in A(\mathcal{F})\}) \tag{3.3.16}$$

Indeed if $x \in \mathcal{F}$ then by definition $x = \sum_{i\in A(\mathcal{F})} \lambda_i e_i$ with $\lambda_i \geq 0, \sum_{i\in A(\mathcal{F})} \lambda_i = 1$. Therefore $Ax = \sum_{i\in A(\mathcal{F})} \lambda_i A^i \in \mathcal{F}_P$. Conversely, if $Ax \in \mathcal{F}_P$ then it cannot be that $x_j \neq 0$ for some $j \notin A(\mathcal{F})$, otherwise $Ax = \lambda_j A^j + \sum_{i\in\{1,\ldots,n\}\setminus\{j\}} \lambda_i A^i$ with

$\lambda_i \geq 0 \ \forall i \in \{1, ..., n\}$, $\lambda_j > 0$. This would mean in particular $Ax \notin \mathcal{F}_P$, contradiction. With this we proved that $\mathcal{F}$ is a face of Δ_{n-1}.
It remains to check that if $\mathcal{A}_P \subset X_P^* \cap \mathcal{F}_P$ has geometric support $\mathcal{F}_P$ then also $A^{-1}(\mathcal{A}_P) \subset X^* \cap \mathcal{F}$ has geometric support $\mathcal{F}$. The $\subset$ relation follows clearly from the definitions, so it only remains to prove that $E(-\nabla f(x)) = \mathcal{F}$ for every $x \in \mathcal{A}$. Equivalently we want to prove

$$(-\nabla f(x), y) = (-\nabla f(x), x) \ \forall \ x \in \mathcal{A}, \ (-\nabla f(x), y) < (-\nabla f(x), x) \ \forall \ y \notin \mathcal{F} \tag{3.3.17}$$

But $-\nabla f(x) = -\nabla f_P(Ax)A$, $\mathcal{A} = A^{-1}\mathcal{A}_P$ so that the first piece can be rewritten as

$$(-\nabla f_P(Ax)A, y) = (-\nabla f_P(Ax)A, x) \ \forall \ y \in A^{-1}(\mathcal{F}_P) \tag{3.3.18}$$

or equivalently

$$\begin{aligned}(-\nabla f_P(Ax), Ay) =&(-\nabla f_P(Ax), Ax) \ \forall \ y \in \mathcal{A}^{-1}(\mathcal{F}_P) \Leftrightarrow (-\nabla f_P(Ax), z) = \\ =&(-\nabla f_P(Ax), Ax) \ \forall \ z \in \mathcal{F}_P\end{aligned} \tag{3.3.19}$$

which is true because by hypotesis $E_P(-\nabla f_P(Ax)) = \mathcal{F}_P$. The second piece of (3.3.17) can be proved analogously.

Remark 3.3.6. We just proved implicitly that $\text{ri}(N(\mathcal{F})) = A^{-1}\text{ri}(N(\mathcal{F}_P))$

We can finally state the main theorem:

Theorem 3.3.7. *Let $\mathcal{P}$ be a finite set of points in $\mathbb{R}^n$, $P = \text{conv}(\mathcal{P})$, $f : P \to \mathbb{R}$ with continuous differential, $\mathcal{A} \subset X^*$ with geometric support $\mathcal{F}$. Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by the AFW or the PFW converging to $\mathcal{A}$ using linesearch for the step size. Then $x_k \in \mathcal{F}$ for k large enough.*

By the previous reasoning we can assume without loss of generality $\mathcal{P} = \{e_i\}_{1\leq i\leq n}$ so that $P = \Delta_{n-1}$. Indeed once we've proved the statement for Δ_{n-1} we can generalize to $\text{conv}(\mathcal{P})$ by conjugation. One key point is that $\{x_k\} \to \mathcal{A}$ also for the transformed sequence in Δ_{n-1} because the transformation $A : \Delta_{n-1} \to \text{conv}(\mathcal{P})$ is surjective.
We need to check that the AFW and the PFW satisfy all the hypotheses of the more abstract theorems proved, and for the PFW additional considerations will be also needed to complete the proof.
In the rest of this section $\mathcal{F} = \{x \in \Delta_{n-1} \mid x_i = 0, \ i \in I^c\}$ will be a face of Δ_{n-1}.

Lemma 3.3.8.

$$\begin{aligned}\text{ri}(\mathcal{F}) &= \{x \in \mathcal{F} \mid x_j > 0 \ \forall \ j \in I\} \\ \text{aff}(\mathcal{F}) &= \{x \in \mathbb{R}^n \mid x_i = 0 \ \forall \ i \in I^c, \ \sum_{j\in I} x_j = 1\}\end{aligned} \tag{3.3.20}$$

Proof. Follows immediately from the definitions. □

Lemma 3.3.9. *There exists a neighborhood U of $\mathcal{A}$ such that for every $x \in U$:*

$$\begin{aligned}\text{argmin}_{i\in\{1,...,n\}}(-\nabla f(x), e_i) &\subseteq I \\ \min_{i\in I^c}(-\nabla f(x), e_i) &> \max_{i\in I}(-\nabla f(x), e_i) \\ \min_{i\in I^c}(-\nabla f(x), e_i - x) &> \max_{i\in I}(-\nabla f(x), x - e_i)\end{aligned} \tag{3.3.21}$$

Proof. We recall that $\mathcal{F} = \{x \in \Delta_{n-1} \mid x_i = 0 \ \forall i \in I^c\}$ is by definition the exposed face by $-\nabla f(x)$ for every $x \in \mathcal{A}$. This means that for every $x \in \mathcal{A}$ we have $(\nabla f(x), e_i) = (\nabla f(x), x)$ for every $i \in I$ and $(-\nabla f(x), e_i) > (-\nabla f(x), x)$ for every $i \in I^c$. The inequalities in (3.3.21) then follow immediately for any $x \in \mathcal{A}$ and by continuity also in a neighborhood of every $x \in \mathcal{A}$. But $\mathcal{A}$ is compact so that the inequalities in (3.3.21) hold in a neighborhood of $\mathcal{A}$. □

Lemma 3.3.10. *For every $x \in U$ defined as in Lemma 3.3.9 if $x \notin \mathcal{F}$ the away direction selected by a FW variant d^{AW} is equal to $x - e_i$ for some $i \in I^c$ and the classic FW direction d^{FW} is equal to $e_i - x$ for some $i \in I$. Moreover, the AFW select d^{AW} as search direction.*

Proof. By definition $d^{AW} = x - e_i$ with $i \in \operatorname{argmax}_{\substack{i \in \{1,\dots,n\} \\ x_i > 0}} (\nabla f(x), e_i)$. But since $x \notin \mathcal{F}$ there exists $i \in I^c$ such that $x_i > 0$, so that by the second equation of (3.3.21) we have $\operatorname{argmax}_{i \in \{1,\dots,n\}} (\nabla f(x), e_i) \subseteq I^c$ for every $x \in U$.
We have also $d^{FW} = e_i - x$ with $i \in \operatorname{argmin}_{i \in \{1,\dots,n\}} (\nabla f(x), e_i) \subseteq I$.
It remains to prove that the AFW select d^{AW} or writing explicitly the selection rule $(d^{AW}, -\nabla f(x)) > (d^{FW}, -\nabla f(x))$ for every $x \in U \setminus \mathcal{F}$. But

$$(d^{AW}, -\nabla f(x)) = (-\nabla f(x), x - e_i)$$

for some $i \in I^c$ and

$$(d^{FW}, -\nabla f(x)) = (-\nabla f(x), e_j - x)$$

for some $j \in I$ so that $(d^{AW}, -\nabla f(x)) > (d^{FW}, -\nabla f(x))$ follows by the third inequality in (3.3.21). □

Lemma 3.3.11. *Under the notation introduced above*

$$\sup_{\substack{y \in \mathcal{F} \\ i \in I^c}} \frac{\|\pi_{\operatorname{lin}(T(\mathcal{F})}(e_i - y)\|}{\|e_i - y\|} = h(\mathcal{F}) < 1 \tag{3.3.22}$$

Proof. Since I^c is finite it suffices to prove the inequality for every $i \in I^c$. Since $\operatorname{lin}(T(\mathcal{F})) = \operatorname{aff}(\mathcal{F}) - \{y\}$ clearly by the characterization in Lemma 3.3.8 $e_i - y \notin \operatorname{lin}(T(\mathcal{F}))$ so that

$$h_i(y) \stackrel{\text{def}}{=} \frac{\|\pi_{\operatorname{lin}(T(\mathcal{F})}(e_i - y)\|}{\|e_i - y\|} < 1$$

for every $y \in \mathcal{F}$. But then since $h_i(y)$ is continuous

$$\sup_{y \in \mathcal{F}} h_i(y) = \max_{y \in \mathcal{F}} h_i(y) < 1 \tag{3.3.23}$$

□

Lemma 3.3.12. *Let $\mathcal{K} = \{k \in \mathbb{N} \mid x_k \notin \mathcal{F}\}$. Then the sequence of search directions $\{d_k\}_{k \in \mathcal{K}}$ generated by the AFW and the PFW is a sequence of identifying directions for $\mathcal{F}$*

Proof. We first prove the lemma for the PFW search directions. If U is the neighborhood of Lemma 3.3.10, then for k large enough $x_k \in U$ so that

$$d_k^{PFW} = d_k^{FW} + d_k^{AW} = e_i - e_j \tag{3.3.24}$$

for some $i \in I$, $j \in I^c$. since for any $x \in \mathrm{ri}(\mathcal{F})$ clearly $-d_k^{PFW}$ is a feasible direction, $d_k^{PFW} \in -T(\mathcal{F})$ whenever (3.3.24) is satisfied and in particular for k large enough. It remains to show that for k large enough $\|\pi_{\mathrm{lin}(T(\mathcal{F}))}(d_k^{PFW})\| < h\|d_k^{PFW}\|$ for some fixed $h < 1$. But even without explicitly computing the projection, by (3.3.24) it follows that d_k^{PFW} takes a finite number of values and is never in $\mathrm{lin}(T(\mathcal{F}))$. Then for some $M \in \mathbb{N}$ there exists h such that

$$\sup_{k \geq M} \frac{\|\pi_{\mathrm{lin}T(\mathcal{F})}(d_k^{PFW})\|}{\|d_k^{PFW}\|} = h < 1 \tag{3.3.25}$$

because $\|d_k^{PFW}\| > \|\pi_{\mathrm{lin}T(\mathcal{F})}(d_k^{PFW})\|$ for every $k \geq M$ and d_k^{PFW} takes a finite number of values.

We now prove the lemma for the AFW search directions. Let $M, \varepsilon > 0$ be such that $x_k \in U \cap B(\mathcal{F}, \varepsilon)$ for every $k \geq M$ with $\mathrm{dist}(e_i, \mathcal{F}) > 2\varepsilon$ for every $i \in I$. Then by Lemma 3.3.10 we have $d_k = d_k^{AW} = x_k - e_i$ for some $i \in I^c$ for every $k \geq M$ such that $x_k \notin \mathcal{F}$. It is now convenient to split the iteration indexes in a family of sets $\{\mathcal{K}_i\}_{i \in I^c}$ defined by

$$\mathcal{K}_i = \{k \geq M \mid d_k^{AW} = x_k - e_i\} \tag{3.3.26}$$

so that

$$\mathcal{K} = \{k \geq M \mid x_k \notin \mathcal{F}\} = \cup_{i \in I^c} \mathcal{K}_i \tag{3.3.27}$$

Fix $i \in I^c$. For every $k \in \mathcal{K}_i$, let $a_k = \pi_{\mathcal{F}}(x_k)$ so that

$$\mathrm{dist}(-T(\mathcal{F}), d_k^{AW}) = \mathrm{dist}(-T(\mathcal{F}), x_k - e_i) \leq \mathrm{dist}(a_k - e_i, x_k - e_i) = \|a_k - x_k\| \tag{3.3.28}$$

where the inequality is justified because $a_k \in \mathcal{F}$ so that $a_k - e_i \in -T(\mathcal{F})$. We can finally write

$$\limsup_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \mathrm{dist}(-T(\mathcal{F}), d_k^{AW}) \leq \limsup_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \|a_k - x_k\| = 0 \tag{3.3.29}$$

Notice that by the definition of M we have $\|d_k^{AW}\| = \|e_i - x_k\| > \varepsilon$, so that

$$\limsup_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \mathrm{dist}(-T(\mathcal{F}), \frac{d_k^{AW}}{\|d_k^{AW}\|}) < \limsup_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \frac{\mathrm{dist}(-T(\mathcal{F}), d_k^{AW})}{\varepsilon} = 0 \tag{3.3.30}$$

We now prove that for every $k \in \mathcal{K}_i$ we also have

$$\frac{\|\pi_{\mathrm{lin}T(\mathcal{F})}(d_k^{AW})\|}{\|d_k^{AW}\|} < h < 1 \tag{3.3.31}$$

for some fixed $h < 1$. First, notice that since $(d_k^{AW})_i < 0$ for every $k \in \mathcal{K}_i$ then $d_k^{AW} \notin \mathrm{lin}(T(\mathcal{F}))$ so that the fraction in (3.3.31) is < 1. We also have

$$\lim_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \frac{\|d_k^{AW}\|}{\|e_i - a_k\|} = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}_i}} \frac{\|\pi_{\mathrm{lin}(T(\mathcal{F}))}(d_k^{AW})\|}{\|\pi_{\mathrm{lin}(T(\mathcal{F}))}(e_i - a_k)\|} = 1 \tag{3.3.32}$$

so that

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}_i}} \frac{\|\pi_{\text{lin}T(\mathcal{F})}(d_k^{AW})\|}{\|d_k^{AW}\|} = \limsup_{\substack{k\to\infty \\ k\in\mathcal{K}_i}} \frac{\|\pi_{\text{lin}(T(\mathcal{F}))}(e_i - a_k)\|}{\|e_i - a_k\|} \leq h(\mathcal{F}) < 1 \tag{3.3.33}$$

where the constant $h(\mathcal{F})$ is the one of Lemma 3.3.11 and does not depend on i. Equation (3.3.33) together with the fact that the fraction in (3.3.31) is always < 1 gives that for some constant k_i

$$\frac{\|\pi_{\text{lin}T(\mathcal{F})}(d_k^{AW})\|}{\|d_k^{AW}\|} \leq k_i < 1 \tag{3.3.34}$$

so that for every $k \geq M$ such that $x_k \notin \mathcal{F}$

$$\frac{\|\pi_{\text{lin}T(\mathcal{F})}(d_k^{AW})\|}{\|d_k^{AW}\|} \leq \max_{i\in I} k_i < 1 \tag{3.3.35}$$

□

We can now prove the main theorem:

Proof of Theorem 3.3.7. Let M be as in Lemma 3.3.12 and $k \geq M$. We first want to prove that if $x_k \in \mathcal{F}$ then $x_{k+1} \in \mathcal{F}$ for both the AFW and the PFW. But this is true because $d_k^{AW} = x - e_i$ for some $e_i \in \mathcal{F}(x) \subseteq \mathcal{F}$, and by Lemma 3.3.9 we have $d_k^{FW} = e_j - x$ for some $e_j \in \mathcal{F}$. Then $x_k + \lambda d_k \in \text{aff}(\mathcal{F})$ for every $\lambda \in \mathbb{R}$ and in particular $x_{k+1} \in \mathcal{F}$.
On the other hand if $\mathcal{K}$ is the set of indexes such that $x_k \notin \mathcal{F}$ we've already proved in lemma 3.3.12 that $\{d_k\}_{k\in\mathcal{K}}$ is a sequence of identifying directions for $\mathcal{A}$. Thus for the AFW the only condition that remains to be proved to apply Proposition 3.3.5 is $x_{k+1} \in \mathcal{F}(x_k)$. Indeed we have $d_k = d_k^{AW} = e_i - x_k$ for some $e_i \in \mathcal{F}(x_k) \setminus \mathcal{F}$ by Lemma 3.3.9 and this concludes the proof for the AFW.
As for the PFW, let $J_k = \{i \in I^c \mid (x_k)_i > 0\}$. Since $d_k^{PFW} = e_i - e_j$ for $i \in I$, $j \in J_k$ and by Lemma 3.3.4 the $x_{k+1} = x_k + \lambda_{\max} d_k^{PFW} = x_k + (x_k)_j d_k^{PFW}$ it follows that $(x_{k+1})_j = 0$, $(x_{k+1})_h = (x_k)_h$ for every $h \notin J_k$ so that $|J_{k+1}| < |J_k|$. It follows that eventually there must be $\bar{k}$ such that $|J_{\bar{k}}| = 0$, or equivalently such that $x_{\bar{k}} \in \mathcal{F}$. This together with the first part implies $x_{\bar{k}+m} \in \mathcal{F}$ for every $m \in \mathbb{N}$. □

Chapter 4

AFW active set complexity

4.1 Introduction and preliminaries

In this chapter we give explicit bounds for the AFW active set complexity on different settings. We mostly analyze applications of the AFW over the simplex, which is not restrictive with respect to the general polytope setting by the affine invariance properties of the AFW (see for instance [28]). In fact every application of the AFW to a polytope can be seen as an application of the AFW to the simplex, with each vertex of the simplex corresponding to one of the atoms generating the polytope.

The key idea in the complexity proofs is that there exists a neighborhood of the set of minimizers for which the AFW at each iteration identifies an active constraint. In particular to bound the active set complexity it is sufficient to control how many iterations it takes for the AFW sequence to enter this neighborhood.

Finite time active set complexity for the AFW on the simplex has been proved recently in [9]. However the proof used additional hypotheses on the curvature of f, which we will not use here, and no explicit bounds were given. Here we use also a slightly different definition of support identification, which for general polytopes can be nicely translated in terms of exposed faces as we will show later in this chapter, subsection 4.6.4.

In the rest of this chapter $f : \Delta_{n-1} \to \mathbb{R}$ will be a function with gradient having Lipschitz constant L and X^* will be the set of minimizers of f. The constant L will also be used as Lipschitz constant for ∇f with respect to the norm $\| \cdot \|_1$. This does not require any additional hypothesis on f since in general $\| \cdot \|_1 \geq \| \cdot \|$ so that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \leq L\|x - y\|_1 \tag{4.1.1}$$

for every $x, y \in \Delta_{n-1}$.

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

$$\text{dist}(A, X) = \inf\{a \in A, x \in X \mid \|a - x\|\} \tag{4.1.2}$$

We define dist_1 in the same way but with respect to $\| \cdot \|_1$. Given a (convex and bounded) polytope P and a linear function c we define the face of P *exposed by c* as

$$\mathcal{F}(c) = \text{argmax}\{cx \mid x \in P\} \tag{4.1.3}$$

It follows from the definition that the face of P exposed by a linear function is always unique and non empty.
We now introduce the multiplier functions, which were recently used in [18] to define an active set strategy for minimization over the 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

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

or in vector form

$$\lambda(x) = \nabla f(x) - (x, \nabla f(x))e \tag{4.1.4}$$

Remarkably, for every $x \in X^*$ these functions coincide with the Lagrangian multipliers of the constraints $x_i \geq 0$.

4.2 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 minimizer at every step the AFW identifies one variable violating the complementarity conditions with respect to the multiplier functions on this minimizer (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 minimizer for this identifying away step to take place.
A lower bound on the size of the non maximal away steps is needed in the following theorem, otherwise of course the steps could be arbitrarily small and there could be no convergence at all.
We use the notation introduced in [34] for the FW direction $d_k^{\mathcal{FW}}$ and the away direction $d_k^{\mathcal{A}}$.

Theorem 4.2.1. *Let x^* be a fixed point in X^*, let*

$$I = \{i \in \{1, ..., n\} \mid \lambda_i(x^*) = 0\}$$

and let $I^c = \{1, ...n\} \setminus I$. Let $\{x_k\}_{k \in \mathbb{N}_0}$ be the sequence of points generated by the AFW,

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

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

$$|J_{k+1}| \leq \max\{0, |J_k| - 1\} \tag{4.2.1}$$

Before proving the main theorem we need to compute the local Lipschitz constant of λ in x^*.

Lemma 4.2.2. *Given $h > 0$, $x_k \in \Delta_{n-1}$ such that $\|x_k - x^*\|_1 \leq h$ let*

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

and assume that $O_k \neq I^c$. *Let* $\delta^k = \max_{i,j\in\{1,\ldots,n\}\setminus O_k} \lambda_i(x^*) - \lambda_j(x^*)$. *For every* $i \in \{1, ..., n\}$*:*

$$|\lambda_i(x^*) - \lambda_i(x_k)| \leq h(L + \frac{\delta^k}{2}) \tag{4.2.2}$$

In this proof we simply show that we have all the hypotheses to apply the technical Lemma 4.6.1 to bound the left hand side of (4.2.2).

Proof. Let $\bar{\lambda} : \Delta_{n-1} \times \mathbb{R}^n \to \mathbb{R}^n$ be defined by $\bar{\lambda}(x, a) = a - (a, x)e$ so that for every $x \in \Delta_{n-1}$

$$\bar{\lambda}(x, \nabla f(x)) = \lambda(x) \tag{4.2.3}$$

and in particular

$$|\lambda_i(x_k) - \lambda_i(x^*)| = |\bar{\lambda}_i(x_k, \nabla f(x_k)) - \bar{\lambda}_i(x^*, \nabla f(x^*))| \tag{4.2.4}$$

for every $i \in \{1, ..., n\}$.
We have $\|x^* - x_k\|_1 \leq h$, $x_i^* = (x_k)_i = 0$ for $i \in O_k$ by hypothesis and

$$\|\nabla f(x^*) - \nabla f(x_k)\|_1 \leq L\|x^* - x_k\| \leq L\|x^* - x_k\|_1 \leq Lh \tag{4.2.5}$$

by the Lipschitz condition. This means that using the notation of Lemma 4.6.1 we have $(x_k, \nabla f(x_k)) \in P_{h,L}^{O_k}(x^*, \nabla f(x^*))$ and by applying the lemma we get the inequality

$$|\bar{\lambda}_i(x^*, \nabla f(x^*)) - \bar{\lambda}_i(x_k, \nabla f(x_k))| \leq h(L + \frac{\delta_{\max}^{O_k}(\nabla f(x^*))}{2})$$

Concatenating this to (4.2.4) we obtain

$$|\lambda_i(x^*) - \lambda_i(y)| \leq h(L + \frac{\delta_{\max}^{O_k}(\nabla f(x^*))}{2})$$

where

$$\delta_{\max}^{O_k}(\nabla f(x^*)) = \max_{i,j\in\{1,\ldots,n\}/O_k} \nabla f_i(x^*) - \nabla f_j(x^*) = \max_{i,j\in\{1,\ldots,n\}/O_k} \lambda_i(x^*) - \lambda_j(x^*) = \delta^k$$

□

We now show a few important relations between the multipliers and the directions selected by the AFW algorithm. Notice that for a fixed x_k the multipliers $\lambda_i(x_k)$ are the values of the linear function $x \to (\nabla f(x_k), x)$ on the vertexes of Δ_{n-1} up to a constant, which in turns are the values controlled by the AFW to select the direction, so the next results should not be surprising.

Lemma 4.2.3. *Let* $H_k = \{i \in \{1, ..., n\} \mid (x_k)_i > 0\}$. *Then*

(a) If $\max\{\lambda_i(x_k) \mid i \in H_k\} > \max\{-\lambda_i(x_k) \mid i \in \{1, ..., n\}\}$, *then the AFW does an away step with* $d_k = d_k^{\mathcal{A}} = x_k - e_i$ *for some* $i \in \operatorname{argmax}\{\lambda_i(x_k) \mid i \in H_k\}$.

(b) For every $i \in \{1, ..., n\}/H_k$ *if* $\lambda_i(x_k) > 0$ *then* $(x_{k+1})_i = (x_k)_i = 0$.

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

$$d_k^{\mathcal{A}} = x_k - e_i \quad \text{for some } i \in \operatorname{argmax}\{(-\nabla f(x_k), x_k - e_i) \mid i \in H_k\} = \operatorname{argmax}\{\lambda_i(x_k) \mid i \in H_k\} \tag{4.2.6}$$

As a consequence of (4.2.6)

$$(-\nabla f(x_k), d_k^{\mathcal{A}}) = \max\{(-\nabla f(x_k), d) \mid d = -e_i + x_k, i \in H_k\} = \max\{\lambda_i(x_k) \mid i \in H_k\} \tag{4.2.7}$$

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

$$\begin{aligned}(-\nabla f(x_k), d_k^{\mathcal{FW}}) &= \max\{(-\nabla f(x_k), d) \mid d = e_i - x_k, i \in \{1, ...n\}\} = \\ &= \max\{-\lambda_i(x_k) \mid i \in \{1, ...n\}\}\end{aligned} \tag{4.2.8}$$

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

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

where we used (4.2.7) and (4.2.8) for the first and the second equality respectively, and the inequality is true by hypothesis.
(b) We will first show that $(d_k)_i = 0$ for every $i \in \{1, ..., n\}/H_k$ such that $\lambda_i(x_k) > 0$. We distinguish two cases.
Case 1: $d_k = d_k^{\mathcal{A}} = x_k - e_j$ for some $j \in H_k$. Since $(x_k)_i = 0$ for every $i \in \{1, ..., n\}/H_k$ we also have $(d_k^{\mathcal{A}})_i = (x_k)_i - (e_j)_i = (e_j)_i = 0$ were the last equality is justified because $j \in H_k$ so that in particular $j \neq i$.
Case 2: $d_k = d_k^{\mathcal{FW}}$. We will assume that the minimization oracle selects a vertex solution which simplifies the proof and in practice is often true. In section 4.6 we prove that this additional assumption is not necessary anyway.
Let $d_k^{\mathcal{FW}} = e_j - x_k$ with

$$j \in \operatorname{argmin}\{(\nabla f(x_k), e_l - x_k) \mid l \in \{1, ..., n\}\}$$

We can now prove that $\lambda_i(x_k) > 0$ implies $i \neq j$:

$$\begin{aligned}&(e_j - x_k, \nabla f(x_k)) = (d_k^{\mathcal{FW}}, \nabla f(x_k)) = \min\{(\nabla f(x_k), e_l - x_k) \mid l \in \{1, ..., n\}\} = \\ &= \min\{(\nabla f(x_k), x - x_k) \mid x \in \Delta_{n-1}\} \leq (\nabla f(x_k), x_k - x_k) = 0 < \lambda_i(x_k) = (e_i - x_k, \nabla f(x_k))\end{aligned} \tag{4.2.9}$$

In particular

$$(d_k^{\mathcal{FW}})_j = (e_j - x_k)_i = (e_j)_i - (x_k)_i = 0 \tag{4.2.10}$$

To finish the proof, just observe that $(x_{k+1})_i = (x_k)_i + \gamma(d_k)_i$ by definition with $(x_k)_i + \gamma(d_k)_i = 0$ for every i such that $(x_k)_i = 0$ and $\lambda_i(x_k) > 0$ because also the second summand in both of the two cases we just examined is 0. $\square$

We can now prove the main theorem. The strategy will be to split $\{1, ..., n\}$ in three subsets I, J_k and $O_k = I^c/J_k$ and use Lemma 4.2.2 to control the variation of the multiplier functions on each of these three subsets. In the proof we examine two possible cases under the assumption of being close enough to a minimum. If $J_k = \emptyset$, which means that the current iteration of the AFW has identified the support of the solution, then we will show that the AFW choses a direction contained in the support so that also $J_{k+1} = \emptyset$.
If $J_k \neq \emptyset$, we will show that in the neighborhood claimed by the theorem the largest multiplier in absolute value is always positive, with index in J_k, and big enough so that the corresponding away step is maximal. This means that the AFW at the iteration $k+1$ identifies a new active variable.

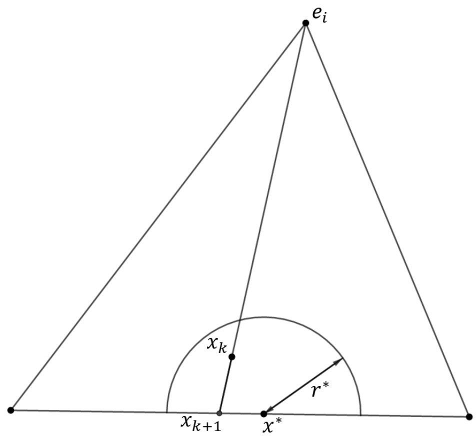


Figure 4.1: Away step identifies one active variable

Proof. If $\lambda(x^*) = 0 \Leftrightarrow I^c = \emptyset$ then there is nothing to prove since $J_k \subset I^c = \emptyset \Rightarrow |J_k| = |J_{k+1}| = 0$.
Otherwise since $I^c \neq \emptyset$ and since by optimality conditions $\lambda_i(x^*) \geq 0$ for every i necessarily $\delta_{\min} > 0$.
As in Lemma 4.2.2, let $O_k = \{i \in I^c \mid (x_k)_i = 0\}$, so that $I^c/O_k = J_k$ and

$$\begin{aligned}\delta^k = \max_{i,j \in \{1,...,n\}/O_k} \lambda_i(x^*) - \lambda_j(x^*) &= \max_{i \in \{1,...,n\}/O_k} \lambda_i(x^*) - \min_{j \in \{1,...,n\}/O_k} \lambda_j(x^*) = \\ &= \max_{i \in J_k \cup I} \lambda_i(x^*) - \min_{j \in J_k \cup I} \lambda_j(x^*) = \max_{i \in J_k \cup I} \lambda_i(x^*)\end{aligned} \tag{4.2.11}$$

where in the last equality we used that $\lambda_j(x^*) \geq 0$ for every j and that $I \neq \emptyset$ so that $\min_{j \in J_k \cup I} \lambda_j(x^*) = 0$. For every $i \in \{1, ..., n\}$, by Lemma 4.2.2

$$\begin{aligned}\lambda_i(x_k) = \lambda_i(x^* + (x_k - x^*)) \geq \lambda_i(x^*) - \|x_k - x^*\|_1 (L + \frac{\delta^k}{2}) > \\ > \lambda_i(x^*) - r_*(L + \frac{\delta^k}{2}) = \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta^k}{2})}{2L + \delta_{\min}}\end{aligned} \tag{4.2.12}$$

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 $i \in I$. Equation (4.2.2) becomes

$$\lambda_i(x_k) > \lambda_i(x^*) - \frac{\delta_{\min} L}{2L + \delta_{\min}} \tag{4.2.13}$$

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

$$\lambda_i(x_k) > \delta_{\min} - \frac{\delta_{\min} L}{2L + \delta_{\min}} > 0 \tag{4.2.14}$$

This means that for every $i \in I^c$ we have $(x_k)_i = 0$ and $\lambda_i(x_k) > 0$, so we can apply part (b) of Lemma 4.2.3 and conclude $(x_{k+1})_i = 0$ for every $i \in I^c$.

Case 2. $|J_k| > 0$. In particular if $i \in \text{argmax}\{i \in J_k \mid \lambda_i(x^*)\}$ we have

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

where we used that $\lambda_j(x^*) = 0 < \lambda_i(x^*)$ for every $j \in I$. Then by the definition (4.2.11) it follows

$$\lambda_i(x^*) = \delta^k$$

so that

$$\lambda_i(x_k) > \lambda_i(x^*) - \frac{\delta_{\min}(L + \frac{\delta^k}{2})}{2L + \delta_{\min}} \geq \frac{\delta^k}{2} \tag{4.2.15}$$

where we used (4.2.12) and that $\delta^k \geq \delta_{\min}$.
We will now show that $d_k = x_k - e_i$ with $i \in J_k$.
For every $i \in I$ since $\lambda_i(x^*) = 0$ again by Lemma 4.2.2

$$\begin{aligned} |\lambda_i(x_k)| = |\lambda_i(x_k) - \lambda_i(x^*)| \leq \|x_k - x^*\|_1 (L + \delta^k/2) < \\ < r_*(L + \delta^k/2) \leq \delta^k/2 \end{aligned} \tag{4.2.16}$$

and for every $i \in I^c$ by (4.2.12)

$$\lambda_i(x_k) > \delta_{\min} - \frac{\delta_{\min}(L + \frac{\delta^k}{2})}{2L + \delta_{\min}} > -\frac{\delta^k}{2} \tag{4.2.17}$$

Then using this together with (4.2.16), (4.2.15) we get $-\lambda_j(x_k) < \delta^k/2 < \lambda_h(x_k)$ for every $j \in \{1, ..., n\}, h \in \text{argmax}\{\lambda_i(x^*) \mid i \in J_k\}$. So the hypothesis of Lemma 4.2.3 is satisfied and $d_k = d_k^{\mathcal{A}} = x_k - e_i$ with $i \in \text{argmax}\{\lambda_i(x_k) \mid i \in H_k\}$. We need to show $i \in J_k$. But $H_k \subseteq I \cup J_k$ and by (4.2.16) if $i \in I$ then $\lambda_i(x_k) < \delta^k/2 < \lambda_j(x_k)$ for every $j \in \text{argmax}\{i \in J_k \mid \lambda_i(x^*)\}$. If $i \in O_k$ then $(x_k)_i = 0$ and $i \notin H_k$. Hence we can conclude $\text{argmax}\{\lambda_i(x_k) \mid i \in H_k\} \subseteq J_k$ and $d_k = x_k - e_i$ with $i \in J_k$. In particular, by (4.2.15) we get

$$\lambda_i(x_k) = \max\{\lambda_j(x_k) \mid j \in J_k\} > \frac{\delta^k}{2} \tag{4.2.18}$$

We now want to show that $\alpha_k = \alpha_{\max}$. Assume by contradiction $\alpha_k < \alpha_{\max}$. Then

$$\alpha_k \geq \frac{(-\nabla f(x_k), d_k)}{L\|d_k\|^2} = \frac{\lambda_i(x_k)}{L\|d_k\|^2} > \frac{\delta_{\min}}{2L\|d_k\|^2} \tag{4.2.19}$$

where in the last inequality we used (4.2.18) together with $\delta^k \geq \delta_{\min}$. Also, by Lemma 4.6.3

$$\begin{aligned} &\|d_k\| = \|e_i - x_k\| \leq \sqrt{2}(e_i - x_k)_i = -\sqrt{2}(d_k)_i \Rightarrow \frac{(d_k)_i}{\|d_k\|^2} \leq -1/2 \\ &(x_k)_i = (x_k - x^*)_i \leq \frac{\|x_k - x^*\|_1}{2} < \frac{r_*}{2} = \frac{\delta_{\min}}{4L + 2\delta_{\min}} \end{aligned} \tag{4.2.20}$$

Finally, combining (4.2.20) with (4.2.19)

$$(x_{k+1})_i = (x_k)_i + (d_k)_i\alpha_k < \frac{r_*}{2} + (d_k)_i \frac{\delta_{\min}}{2L\|d_k\|^2} \leq$$
$$\leq \frac{\delta_{\min}}{4L + 2\delta_{\min}} - \frac{\delta_{\min}}{4L} < 0$$

where we used (4.2.19) to bound α in the first inequality, (4.2.20) to bound $(x_k)_i$ and $\frac{(d_k)_i}{\|d_k\|^2}$. Hence $(x_{k+1})_i < 0$, contradiction. □

4.3 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 solution x^* we do not mean the set $A(x^*) = \{i \in \{1, ..., n\} \mid (x^*)_i = 0\}\}$ but the set $I^c(x^*) = \{i \in \{1, ..., n\} \mid \lambda_i(x^*) > 0\}$. In general $I^c(x^*) \subset A(x^*)$ by complementarity conditions and the two sets coincide under strict complementarity conditions. The face $\mathcal{F}$ of Δ_{n-1} defined by the constraints with indexes in $I^c(x^*)$ still 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 $\{x_k\}_{k\in\mathbb{N}}$ converges to a subset A of X^* for which I^c is constant. This motivates the following definition:

Definition 4.3.1. Given a compact subset A of X^* we will say that the multiplier function λ has the support identification property for A if there exists

$$I^c(A, \lambda) \subset \{1, ..., n\}$$

such that for every $x \in A$ the support of the multiplier function is $I^c(A, \lambda)$. Under these conditions we define

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

The geometrical interpretation of the above definition is the following: for every point in the subset A the negative gradient $-\nabla f(x^*)$ exposes the same face. This is trivially true if A is a singleton, and it is also true if for instance A is contained in the relative interior of a face of Δ_{n-1} and strict complementarity conditions hold for every point in this face.
Notice that by the compactness of A we always have $\delta_{\min}(A, \lambda) > 0$. We can finally give a rigorous definition of what it means to solve the active set problem:

Definition 4.3.2. Consider an instance of the AFW generating a sequence $\{x_k\}_{k\in\mathbb{N}}$ converging to a subset A of X^* for which λ has the support identification property. We will say that this instance solve the active set problem in M steps if $(x_k)_i = 0$ for every $i \in I^c(A, \lambda)$, $k \geq M$.

We can now apply Lemma 4.2.1 to show that once a sequence is definitely close enough to a set for which λ has the support identification property the AFW identifies the active set in at most $|I^c|$ steps.

Theorem 4.3.3. *Let X^* be the set of minimizers of a function $f : \Delta_{n-1} \to \mathbb{R}$ with ∇f having Lipschitz constant L. Let $\{x_k\}$ be a sequence generated by the AFW applied to f on the simplex, and assume that there exists a compact subset A of X such that $\mathrm{dist}(x_k, A) \to 0$ and for every $x \in A$ the support of the multiplier function is I^c. Then there exists M such that $(x_k)_i = 0$ for every $k \geq M$, $i \in I^c(A, \lambda)$.*

Proof. Since λ has the support identification property for A we can set $\delta_{\min} = \delta_{\min}(A, \lambda) > 0$ and $I^c = I^c(A, \lambda)$ to simplify notations. Let $\bar{k}$ be such that $\mathrm{dist}_1(x_k, A) < \frac{\delta_{\min}}{2L+\delta_{\min}} = r_*$ for every $k \geq \bar{k}$, and let $J_k = \{i \in I^c \mid (x_k)_i > 0\}$.
Then every $k \geq \bar{k}$ there exists $y^* \in A$ with $\|x_k - y^*\|_1 < r_*$. But since by hypothesis for every $y^* \in A$ the support of the multiplier function is I^c with $\delta_{\min} \leq \lambda_i(y^*)$ for every $i \in I^c$, we can apply Theorem 4.2.1 with y^* as fixed point and obtain that $J_{k+1} \leq \max(0, J_k - 1)$. This means that it takes at most $|J_{\bar{k}}| \leq |I^c|$ steps for all the variables with indexes in $|I^c|$ to be 0. To conclude, again by (4.2.1) since $|J_{\bar{k}+|I^c|}| = 0$ by induction $|J_m| = 0$ for every $M \geq \bar{k} + |I^c|$. □

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

Proposition 4.3.4. *Under the hypotheses of Proposition 4.3.3 the active set complexity is at most*

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

where $r^* = \frac{\delta_{\min}}{2L+\delta_{\min}}$.

Finally, under some assumptions on the set of minimizers X^* and on the step sizes, we can prove finite time active set identification. This theorem is a consequence of the local convergence properties we just proved combined with a general convergence theorem that we prove in the appendix. In subsection 4.6.3 we discuss the hypotheses on X^* and the step sizes.

Theorem 4.3.5. *Assume that $X^* = \bigcup_{i=1}^{C} A_i$ where $\{A_i\}$ is a family of compact and disjoint sets, and assume that for each of these sets λ has the support identification property. Let $\{x_k\}_{k\in\mathbb{N}}$ be the sequence generated by the AFW with step sizes satisfying $\alpha_k \leq 2(\nabla f(x_k), d_k)/\|d_k\|^2 L$ or more in general*

$$x_k \in \mathrm{argmax}\{f(x) \mid x \in \mathrm{conv}(x_k, x_{k+1})\} \tag{4.3.1}$$

If $f(x_k) \to f^*$ then there exists i such that

$$\mathrm{dist}_1(x_k, A_i) \to 0 \tag{4.3.2}$$

Moreover, $(x_k)_i = 0$ for every $i \in I^c(A_i\lambda)$,

$$i \geq \bar{k} + |I^c(A_i, \lambda)| \tag{4.3.3}$$

where $\bar{k}$ is the minimum such that

$$\mathrm{dist}_1(x_k, A_i) \leq \frac{\delta_{\min}(A_i, \lambda)}{2L + \delta_{\min}(A_i, \lambda)}$$

Proof. Thanks to the condition on the step sizes by Lemma 4.6.5 the AFW satisfies the condition

$$x_k \in \operatorname{argmax}\{f(x) \mid x \in \operatorname{conv}(x_k, x_{k+1})\} \tag{4.3.4}$$

In particular we have all the hypotheses to apply 4.6.4 and obtain that there exists A_i such that $\operatorname{dist}_1(x_k, A_i) \to 0$. The active set complexity bound (4.3.3) follows the immediately from 4.3.4.

□

As an example of a more concrete application of Theorem 4.2.1 we prove an active set complexity result for strongly convex functions on the simplex. We will actually use a slightly weaker hypothesis: f is convex and has a unique minimum x^* on the $n-1$ dimensional simplex Δ_{n-1} such that

$$f(x) \geq \frac{u_1}{2}\|x - x^*\|_1^2 \tag{4.3.5}$$

for every x on Δ_{n-1}.

Corollary 4.3.6. *Let $\{x_k\}_{k\in\mathbb{N}_0}$ be the sequence of points generated by the AFW, $h_k = f(x_k) - f_*$. Let $q < 1$ be such that $h_k \leq q^k h_0$. Under the same hypotheses of Theorem 4.2.1, if also the error bound condition* (4.3.5) *holds, then the active set complexity is*

$$\max(0, \frac{\ln(h_0) - \ln(ur_*^2/2)}{\ln(1/\mathrm{q})}) + |I^c|$$

Proof. Notice that by the linear convergence rate $h_k \leq q^k h_0$ the number of steps that it takes to reach the condition

$$h_k \leq \frac{u_1}{2} r_*^2 \tag{4.3.6}$$

is at most

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

We claim that if condition (4.3.6) holds then it takes at most $|I^c|$ steps for the sequence to be definitely in the active set.
Indeed if $h_k \leq \frac{u}{2} r_*^2$ then necessarily $x_k \in B(x^*, r_*)$ by (4.3.5) and by monotonicity of the bound we then have $x_{k+h} \in B(x^*, r_*)$ for every $h \geq 0$. Once the sequence is definitely in $B(x^*, r_*)$ by (4.2.1) it takes at most $|J_{\bar{k}}| \leq |I^c|$ steps for all the variables with indexes in $|I^c|$ to be 0. To conclude, again by (4.2.1) since $|J_{\bar{k}+|I^c|}| = 0$ by induction $|J_m| = 0$ for every $m \geq \bar{k} + |I^c|$.

□

Remark 4.3.7. In the above proof we did not use Theorem 4.3.5, which would require additionally the sequence $\{f(x_i)\}$ to be decreasing. It is anyway not difficult to see that the hypothesis

$$x_k \in \operatorname{argmax}\{f(x) \mid x \in \operatorname{conv}(x_k, x_{k+1})\} \tag{4.3.7}$$

is only used to show that the sequence $\{x_k\}$ does not escape from the connected components of sublevel sets, which in the convex case is obvious since every sublevel set is connected.

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

4.4 Active set complexity for non convex objectives

In this section we give a more explicit convergence bound for the general non convex case. A fundamental element in our analysis will be the FW gap function $g : \Delta_{n-1} \to \mathbb{R}$ defined as

$$g(x) = (-\nabla f(x), x) - \min\{(-\nabla f(x), y) \mid y \in \Delta_{n-1}\} \ . \tag{4.4.1}$$

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

$$g(x_k) = (-\nabla f(x_k), d_k^{\mathcal{FW}}) \ . \tag{4.4.2}$$

This FW gap function was used in [32] to analyze the convergence rate of the classic FW algorithm for non convex functions. In particular, a convergence rate of $O(\frac{1}{\sqrt{k}})$ was proved for

$$g_k^* = \min_{0 \leq i \leq k} g(x_i) \ . \tag{4.4.3}$$

The key insight of [32] is that to prove a convergence rate for this sequence g_k^* one can extend in a straightforward way the techniques used in the convex case to the non convex one. This does not appear to be true if one still tries to prove a convergence rate for the sequences $\{f(x_k)\}_{k\in\mathbb{N}}$ or $\{\nabla f(x_k)\}_{k\in\mathbb{N}}$. Following this insight we mostly repeat the steps used to compute the convergence rate of the AFW in the (strongly) convex case (see for instance [39]) to prove a convergence rate for $\{g_k^*\}_{k\in\mathbb{N}_0}$ in the non convex one.
In the rest of this section we assume that the AFW starts from a vertex of the simplex. This is not restrictive because otherwise by affine invariance one can apply the same theorems to the AFW starting from e_{n+1} for $\tilde{f} : \Delta_n \to \mathbb{R}$ satisfying

$$\tilde{f}(x) = f((x_1, ..., x_n) + x_{n+1}p) \tag{4.4.4}$$

where $p \in \Delta_{n-1}$ is the desired starting point. We will discuss more in detail the invariance of the AFW under affine transformations in section 4.6.4.

Theorem 4.4.1. *Let $f \in C^1(\Delta_{n-1}, \mathbb{R})$ be with $L-$ Lipschitz differential. Let $f^* = \min_{x\in\Delta_{n-1}} f(x)$, and let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by the AFW algorithm applied to f on Ω with step size*

$$\alpha_k = \min(\alpha_k^{\max}, \frac{1}{L\|d_k\|^2}(-\nabla f(x_k), d_k)) \ . \tag{4.4.5}$$

Assume that the linear minimization oracle always selects a vertex solution, and that the algorithm starts from a vertex. Then for every $T \in \mathbb{N}$

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

Proof. Let $r_k = -\nabla f(x_k)$, let $S_k = \{i \in \{1, ..., n\} \mid (x_k)_i \neq 0\}$ and $g_k = g(x_k)$. We distinguish 3 cases.
Case 1. $\alpha_k < \alpha_k^{\max}$. Then by the standard descent lemma (see [6], Proposition 6.1.2)

$$f(x_k + \alpha d_k) \leq f(x_k) + \alpha(\nabla f(x_k), d_k) + \frac{\alpha^2 L}{2} \|d_k\|^2 \ . \tag{4.4.6}$$

Minimizing the right hand side with respect to α we have that for $\alpha = \alpha_k$

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \leq f(x_k) - \frac{1}{2L\|d_k\|^2}(r_k, d_k)^2 \tag{4.4.7}$$

which rearranging becomes

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L\|d_k\|^2}(r_k, d_k)^2 \geq \frac{1}{2L\|d_k\|^2} g_k^2 \geq \frac{g_k^2}{4L} \tag{4.4.8}$$

where we used $(r_k, d_k) \geq (r_k, d_k^{\mathcal{FW}}) = g_k$ 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: $\alpha_k = \alpha_k^{\max} = 1, d_k = d_k^{\mathcal{FW}}$. Again by the standard descent lemma applied to f with center x_k and $\alpha = 1$

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

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

$$\frac{(-\nabla f(x_k), d_k)}{\|d_k\|^2 L} \geq 1 \Rightarrow -L\|d_k\|^2 \geq (\nabla f(x_k), d_k) \tag{4.4.9}$$

so that

$$f(x_k) - f(x_{k+1}) \geq (-\nabla f(x_k), d_k) - \frac{L}{2}\|d_k\|^2 \geq -\frac{1}{2}(\nabla f(x_k), d_k) = \frac{1}{2} g_T \ . \tag{4.4.10}$$

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

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

with $\alpha_k = \alpha_k^{\max} = \frac{(x_k)_i}{1-(x_k)_i}$. Therefore $(x_{k+1})_j = 0$ for $j \in \{1, ..., n\} \setminus S_k \cup \{i\}$ and $(x_{k+1})_j \neq 0$ for $j \in S_k \setminus \{i\}$. In particular $|S_{k+1}| = |S_k| - 1$. For $i = 1, 2, 3$ let now $n_i(T)$ be the number of Case i steps done in the first T iterations of the AFW. We have by induction on the recurrence relation we proved for $|S_k|$

$$|S_T| - |S_0| \leq n_1(T) + n_2(T) - n_3(T) \tag{4.4.11}$$

for every $T \in \mathbb{N}$.
Since $n_3(T) = T - n_1(T) - n_2(T)$ from (4.4.11) we get

$$n_1(T) + n_2(T) \geq \frac{T + |S_T| - |S_0|}{2} \geq \frac{T}{2} \tag{4.4.12}$$

where we used $|S_0| = 1 \leq |S_T|$ Let now C_i^T be the set of indexes up to $T-1$ corresponding to Case i steps for $i \in \{1,2,3\}$, which satisfies $|C_i^T| = n_i(T)$. We have by summing (4.4.8) and (4.4.10) for the indexes in C_1^T and C_2^T respectively

$$\sum_{k \in C_1^T} f(x_{k+1}) - f(x_k) + \sum_{k \in C_2^T} f(x_{k+1}) - f(x_k) \geq \sum_{k \in C_1^T} \frac{g_k^2}{4L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \tag{4.4.13}$$

We now lower bound the RHS of (4.4.13) in terms of g_T^*

$$\begin{aligned} &\sum_{k \in C_1^T} \frac{g_k^2}{4L} + \sum_{k \in C_2^T} \frac{1}{2} g_k \geq |C_1^T| \min_{k \in C_1^T} \frac{g_k^2}{4L} + |C_2^T| \min_{k \in C_2^T} \frac{g_k}{2} \geq \\ \geq &(|C_1^T| + |C_2^T|) \min(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2}) = (n_1(T) + n_2(T)) \min(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2}) \geq \\ \geq &\frac{T}{2} \min(\frac{(g_T^*)^2}{4L}, \frac{g_T^*}{2}) = \frac{T}{2} \min(\frac{g_T^*}{2}, \frac{(g_T^*)^2}{4L}) \ . \end{aligned} \tag{4.4.14}$$

Since the LHS of (4.4.13) can clearly be upper bounded by $f(x_0) - f^*$ we have

$$f(x_0) - f^* \geq \frac{T}{2} \min(\frac{g_T^*}{2}, \frac{(g_T^*)^2}{4L}) \ . \tag{4.4.15}$$

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

$$g_T^* \leq \frac{4(f(x_0) - f^*)}{T} \tag{4.4.16}$$

and otherwise

$$g_T^* \leq \sqrt{\frac{8L(f(x_0) - f^*)}{T}} \ . \tag{4.4.17}$$

The thesis follows taking the max in the system formed by (4.4.16) and (4.4.17). $\square$

In the rest of this section we will use the notation introduced in Theorem 4.3.5. Before stating the active set complexity bound result, we need to introduce a few new elements. Let

$$r_i = \min(\frac{\delta_{\min}(A_i, \lambda)}{2L + \delta_{\min}(A_i, \lambda)}, \frac{\text{dist}_1(A_i, X^* \setminus A_i)}{2}) \tag{4.4.18}$$

and

$$X_\lambda^* = \bigcup_{i \in \{1, \dots, C\}} B_1(r_i, A_i) \ . \tag{4.4.19}$$

It follows immediately from the definition that the connected components of X_λ^* are $B_1(r_i, A_i)$ for $1 \leq i \leq C$. Let $m = \min\{x \in \partial X_\lambda^* \mid f(x)\}$ so that in particular $m > f^*$ because $X^* \subset (X_\lambda^*)^\circ$. Finally, let

$$\tau = \min\{g(x) \mid x \in f^{-1}([m, +\infty))\} \tag{4.4.20}$$

Theorem 4.4.2. *Assume that in addition to the hypotheses of Theorem 4.3.5 we also have*

$$\alpha_k = \min(\frac{(-\nabla f(x_k), d_k)}{\|d_k\|^2 L}, \alpha_k^{\max}) \ . \tag{4.4.21}$$

Then the constant $\bar{k}$ *appearing in the statement of Theorem* (4.3.5) *satisfies*

$$\bar{k} \le \max(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}) + 1 \ . \tag{4.4.22}$$

Proof. We have all the hypotheses to apply the bound given in Theorem 4.4.1 for g_k^*.

$$g_k^* \le \max(\sqrt{\frac{8L(f(x_0) - f^*)}{k}}, \frac{4(f(x_0) - f^*)}{k}) \tag{4.4.23}$$

It is straightforward to check that if

$$\bar{h} = \lceil \max(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}) \rceil \tag{4.4.24}$$

then

$$g_{\bar{h}}^* < \tau \ . \tag{4.4.25}$$

So that in particular $g(x_k) < \tau$ for some $k \le \bar{h}$. Hence, by the definition of τ we get $f(x_k) < m$ and also $f(x_{\bar{h}}) < m$ given the monotonicity of the AFW with step sizes given by (4.4.21). We claim that $x_h \in X_\lambda^*$ for every $h \ge \bar{h}$. Indeed otherwise since $x_k \to X^* \subset X_\lambda^*$ there would be $h' \ge \bar{h}$ such that $x_{h'} \notin X_\lambda^*$ but $x_{h'+1} \in X_\lambda^*$. But as a consequence we would have on the one hand $m > f(x_{h'})$ with

$$x_{h'} \in \text{argmax}\{f(x) \mid x \in \text{conv}(x_{h'}, x_{h'+1})\}$$

and on the other hand $f(y) \ge m$ for $y \in \text{conv}(x_{h'}, x_{h'+1}) \cap \partial X_\lambda^* \neq \emptyset$, a contradiction. With the same argument we can prove that for every $h \ge \bar{h}$ the point x_h is in the same connected component of X_λ^*, or in other words there exists i such that

$$\text{dist}_1(A_i, x_h) \le r_i \le \frac{\delta_{\min}(A_i, \lambda)}{2L + \delta_{\min}(A_i, \lambda)} \tag{4.4.26}$$

for every $h \ge \bar{h}$. Of course this i must then coincide with the one in the statement of Theorem (4.3.5). The thesis follows immediately from the definition of $\bar{k}$. □

Combining this result with Theorem 4.3.5 we have the following more explicit estimate for the AFW active set complexity in the non convex case.

Corollary 4.4.3. *Under the hypotheses of Theorem 4.4.2 the active set identification complexity is at most*

$$n + \max(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}) + 1 \ . \tag{4.4.27}$$

Proof. We have by Theorem 4.3.5 that the active set complexity is

$$\mathcal{C}(f) \le \bar{k} + |I^c(A_i, \lambda)| \tag{4.4.28}$$

where i is the index such that $x_k \to A_i$. Then

$$\mathcal{C}(f) \le \bar{k} + |I^c(A_i, \lambda)| \le n + \max(\frac{4(f(x_0) - f^*)}{\tau}, \frac{8L(f(x_0) - f^*)}{\tau^2}) + 1 \tag{4.4.29}$$

where we used Theorem 4.4.2 in the second inequality. □

4.5 Conclusions

We proved general results for the AFW finite time active set convergence problem, giving explicit bounds on the number of steps necessary to identify the support of a solution. As an example of application of these results we computed the active set complexity for strongly convex functions. Possible expansions of these results would be to adapt them for other FW variants and compute the active set complexity for non strongly convex functions applying the known convergence results. It also remains to be seen if these identification properties of the AFW can be extended to problems with non linear constraints.

4.6 Technical proofs

4.6.1 Technical inequalities

In the following technical lemma we define a linear programming problem related to the variation of the multiplier functions and compute an upper bound to its optimal value. Before defining the problem we need to introduce some notation. Let $\bar{\lambda} : \Delta_{n-1} \times \mathbb{R}^n \to \mathbb{R}^n$ be defined by $\bar{\lambda}(x, a) = a - (a, x)e$, so that if λ is the vector of multiplier functions for f then $\lambda(x) = \bar{\lambda}(x, \nabla f(x))$. Let $x \in \Delta_{n-1}$, $a \in \mathbb{R}^n$, $h, L > 0$, $O \subset \{1, ..., n\}$ and let

$$P^O_{h,L}(x, a) = \{(y, b) \in \Delta_{n-1} \times \mathbb{R}^n \mid \|y - x\|_1 \leq h,\ \|b - a\|_1 \leq Lh,\ y_i = x_i\ \forall\ i \in O\}$$

Finally, let $\delta^O_{\max}(a) = \max_{i,j \in \{1,...,n\}/O} a_i - a_j$. Then:

Lemma 4.6.1. *Let $m \in \{1, ..., n\}$ and*

$$\begin{aligned} z = \max\ & |\bar{\lambda}_m(y, b) - \bar{\lambda}_m(x, a)| \\ & (y, b) \in P^O_{h,L}(x, a) \end{aligned} \tag{4.6.1}$$

Then $z \leq (L + \delta^O_{\max}(a)/2)h$.

Notice that in principle one could solve two linear programming problems with the same feasible region of (4.6.1) to compute the maximum and the minimum of the objective function without absolute value. However, the computations seem rather complex, so that instead we will just prove the upper bound on z by splitting the objective function in two summands much easier to bound individually. Our estimate turns out to be the actual optimal value for some instances of the problem, as we show in the remark after the proof.

Proof. By the definition of $\bar{\lambda}_m$ and the triangular inequality

$$|\bar{\lambda}_m(y, b) - \bar{\lambda}_m(x, a)| = |b_m - a_m + (a, x - y) + (a - b, y)| \leq |b_m - a_m + (a - b, y)| + |(a, x - y)| \tag{4.6.2}$$

Let $t_i = -y_i$ for $i \in \{1, ..., n\}/\{m\}$, $t_m = 1 - y_m$. Since $y \in \Delta_{n-1}$ and in particular $0 \leq y_i \leq 1$ we have $|t_i| \leq 1$ for every $i \in \{1, ..., n\}$. This implies that

$$|b_m - a_m + (a - b, y)| = |(b - a, t)| \leq \|b - a\|_1 \|t\|_\infty \leq Lh \tag{4.6.3}$$

where the last inequality is justified by the Holder inequality with exponents $1, \infty$ and holds for every $(y,b) \in P^O_{h,l}(x,a)$.
We now bound the second piece. Let

$$k_i = \max\{0, (x-y)_i\},\ l_i = \max\{0, -(x-y)_i\} \tag{4.6.4}$$

$\sum_{i\in\{1,...,n\}} x_i = \sum_{i\in\{1,...,n\}} y_i = 1$ since $x, y \in \Delta_{n-1}$ so that

$$\sum_{i\in\{1,...,n\}} (x-y)_i = \sum_{i\in\{1,...,n\}} k_i - l_i = 0 \Rightarrow \sum_{i\in\{1,...,n\}} k_i = \sum_{i\in\{1,...,n\}} l_i \tag{4.6.5}$$

Moreover, $2\sum_{i\in\{1,...,n\}} k_i = 2\sum_{i\in\{1,...,n\}} l_i = \sum_{i\in\{1,...,n\}} k_i + l_i = \sum_{i\in\{1,...,n\}} |x_i - y_i| \leq h$ so that

$$\sum_{i\in\{1,...,n\}} k_i = \sum_{i\in\{1,...,n\}} l_i \overset{\text{def}}{=} h'/2 \leq h/2 \tag{4.6.6}$$

Let $a_m = \min_{i\in\{1,...,n\}/O} a_i$, $a_M = \max_{i\in\{1,...,n\}/O} a_i$. By definition $k_i, l_i \geq 0$ for every $i \in \{1, ..., n\}$ so that

$$\frac{h'}{2} a_m = a_m \sum_{i\in\{1,...,n\}/O} k_i \leq \sum_{i\in\{1,...,n\}/O} k_i a_i = \sum_{i\in\{1,...,n\}} k_i a_i \tag{4.6.7}$$

where in the last equality we used that by hypothesis $k_i = l_i = 0\ \forall\ i \in O$. Also

$$\sum_{i\in\{1,...,n\}} k_i a_i = \sum_{i\in\{1,...,n\}/O} k_i a_i \leq a_M \sum k_i = \frac{h'}{2} a_M \tag{4.6.8}$$

Reasoning analogously for l_i we get

$$(a,l) \in [\frac{h'}{2}a_m, \frac{h'}{2}a_M],\ (a,k) \in [\frac{h'}{2}a_m, \frac{h'}{2}a_M] \tag{4.6.9}$$

We can finally bound the second piece of (4.6.2)

$$|(a, x-y)| = |\sum_{i\in\{1,...,n\}} a_i(k_i - l_i)| = |(a,k) - (a,l)| \leq \frac{h'}{2}(a_M - a_m) \leq \frac{h}{2}(a_M - a_m) = \frac{h}{2}\delta^O_{\max}(a) \tag{4.6.10}$$

for every $y \in \Delta_{n-1}$.
To conclude, by (4.6.2)

$$\begin{aligned} z &\leq \max\{|b_m - a_m + (a-b,y)| \mid (y,b) \in P^O_{h,L}(x,a)\} + \max\{|(a, x-y)| \mid (y,b) \in P^O_{h,L}(x,a)\} \leq \\ &\leq \max\{|b_m - a_m + (a-b,y)| \mid (y,b) \in P^O_{h,L}(x,a)\} + \max\{|(a,x-y)| \mid y \in \Delta_{n-1}\} \end{aligned} \tag{4.6.11}$$

and bounding the two summands with (4.6.3) and (4.6.10) respectively

$$\max\{|b_m - a_m + (a-b,y)| \mid (y,b) \in P^O_{h,L}(x,a)\} + \max\{|(a,x-y)| \mid y \in \Delta_{n-1}\} \leq (L + \frac{\delta^O_{\max}(a)}{2})h \tag{4.6.12}$$

□

Remark 4.6.2. There are many simpler ways to bound the variation of the multipliers function, so that it is important to underline that the bound we gave on this linear problem can be optimal. We provide a set of vectors for which our estimate coincide with the actual optimal value of the problem. Let x, a, m be such that for some $i, j \neq m$ we have $\delta^O_{\max}(a) = a_i - a_j$ and $x_m = 0$. Let $y_i = x_i + h/2$, $y_j = x_j - h/2$ and $y_{\{1,...,n\}\setminus\{i,j\}} = x_{\{1,...,n\}\setminus\{i,j\}}$. Finally, let $b_{\{1,...,n\}\setminus\{m\}} = a_{\{1,...,n\}\setminus\{m\}}$, $b_m = a_m - Lh$. It is easy to check that if $x_i \leq 1 - h/2$, $x_j \geq h/2$ then (y, b) is feasible for problem (4.6.1) and that $|\bar{\lambda}_m(y, b) - \bar{\lambda}_m(x, a)| = (L + \delta^O_{\max}(a)/2)h$ coincides with the upper bound we proved.

A few elementary properties of the simplex Δ_{n-1} which will allow us to relate different norms restricted to Δ_{n-1} will be useful to show that the AFW away steps are long enough to be maximal with respect to the boundary conditions.

Lemma 4.6.3. *Given $x, y \in \Delta_{n-1}$, $i \in \{1, ..., n\}$:*

1. $\|e_i - x\| \leq \sqrt{2}(e_i - x)_i$;
2. $(y - x)_i \leq \|y - x\|_1/2$

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

$$\|e_i - x\| = (\sum_{j \neq i} x_j^2 + (e_i - x)_i^2)^{\frac{1}{2}} \leq ((\sum_{j \neq i} x_j)^2 + (1 - x_i)^2)^{\frac{1}{2}} = \sqrt{2}(\sum_{j \neq i} x_i) = \sqrt{2}(e_i - x)_i \tag{4.6.13}$$

2. Since $\sum_{i \in \{1,...,n\}} x_i = \sum_{i \in \{1,...,n\}} y_i$ so that $\sum (x - y)_i = 0$ we have

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

and as a consequence

$$\|y - x\|_1 = \sum_{j \in \{1,...,n\}} |(y - x)_j| \geq (y - x)_i + \sum_{j \neq i} (x - y)_j = 2(y - x)_i \tag{4.6.14}$$

□

4.6.2 Oracles that do not guarantee a vertex solution

We now prove Lemma 4.2.3 without assuming that the minimization oracle finds a vertex solution.

Proof. We omit the first part of the proof which has no significant differences with the one of Lemma 4.2.3 and start from case 2 of point b).
We first prove that the smallest multiplier is at most 0:

$$\begin{aligned} &\min\{\lambda_i(x_k) \mid i \in \{1, ..., n\}\} = \min\{(\nabla f(x_k), e_i - x_k) \mid i \in \{1, ..., n\}\} = \\ &= \min\{(\nabla f(x_k), x - x_k) \mid x \in \Delta_{n-1}\} \leq (\nabla f(x_k), x_k - x_k) = 0 \end{aligned} \tag{4.6.15}$$

Let $d_k = d_k^{\mathcal{FW}} = x_f - x_k$ with

$$x_f \in \text{argmin}\{(\nabla f(x_k), x - x_k) \mid x \in \Delta_{n-1}\} = \text{argmin}\{(\nabla f(x_k), x) \mid x \in \Delta_{n-1}\} \stackrel{\text{def}}{=} C_k$$

Since Δ_{n-1} is a polytope with vertexes $\{e_i\}_{1\leq i\leq n}$ we have

$$C_k = \operatorname{conv}(\{e_i \mid e_i \in C_k\}) \tag{4.6.16}$$

where $e_i \in C_k$ if and only if

$$\begin{aligned}\nabla f_i(x_k) &= \min\{(\nabla f(x_k), e_j) \mid j \in \{1, ..., n\}\} = \min\{\lambda_j(x_k) + (x_k, \nabla f(x_k)) \mid j \in \{1, ..., n\}\} = \\ &= \min\{\lambda_j(x_k) \mid j \in \{1, ..., n\}\} + (x_k, \nabla f(x_k))\end{aligned} \tag{4.6.17}$$

and bringing $(x_k, \nabla f(x_k))$ on the left hand side the condition becomes $\lambda_i(x_k) = \min\{\lambda_j(x_k) \mid j \in \{1, ..., n\}\}$. In other words

$$e_i \in C_k \Leftrightarrow \lambda_i(x_k) \in \operatorname{argmin}\{\lambda_j(x_k) \mid j \in \{1, ..., n\}\} \stackrel{\text{def}}{=} A_k$$

Since $x_f \in C_k = \operatorname{conv}\{e_i \mid i \in A_k\}$ we have $x_f = \sum_{i\in A_k} \bar{\alpha}_i e_i$ where $\bar{\alpha}_i \geq 0$, $\sum_{i\in A_k} \bar{\alpha}_i = 1$. As a consequence $(x_f)_j = 0$ for every j such that $\lambda_j(x_k) > 0$, because $\lambda_j(x_k) > 0 \geq \min\{\lambda_i(x_k) \mid i \in \{1, ..., n\}\}$ by (4.6.15), hence $j \notin A_k$. If additionally $(x_k)_j = 0$,

$$(d_k^{\mathcal{FW}})_j = (x_f - x_k)_j = (x_f)_j - (x_k)_j = 0 \tag{4.6.18}$$

so also case 2 is proved.
The conclusion follows as in Lemma 4.2.3. □

4.6.3 Convergence lemma.

The following is a very general and straightforward lemma which ensures that if the set of minimizers of f can be split in a family of disjoint and compact sets then any minimizing sequence with a certain descent property converge to one of these set. The property is

$$x_k \in \operatorname{argmax}\{f(x) \mid x \in \operatorname{conv}(x_k, x_{k+1})\} \tag{4.6.19}$$

and it is obviously stronger than the usual monotonicity. However, if f is convex, this property is equivalent to $f(x_k) \geq f(x_{k+1})$. Indeed given

$$x \in \operatorname{conv}(x_k, x_{k+1}) = \{\lambda x_k + (1-\lambda)x_{k+1} \mid \lambda \in [0, 1]\}$$

for $\lambda \in [0, 1]$ we have

$$f(\lambda x_k + (1-\lambda)x_{k+1}) \leq \lambda f(x_k) + (1-\lambda)f(x_{k+1}) \leq f(x_k)$$

if $f(x_k) \leq f(x_{k+1})$.

Lemma 4.6.4. *Assume that* $X^* = \bigcup_{i=1}^{C} A_i$ *where the* $\{A_i\}_{1\leq i\leq C}$ *is a family of compact and disjoint sets. Assume that* $\{x_k\}$ *is a sequence in* Δ_{n-1} *with the property* (4.6.19). *Then if* $f(x_k) \to f^*$ *there exists* i *such that* $\operatorname{dist}(x_k, A_i) \to 0$.

In the proof we use that as a consequence of (4.6.19) the sequence x_k can not escape from connected components of sublevel sets, so that when it falls into a connected component of a sublevel set close to a certain A_i it cannot reach the other components of X^*.
We define

$$B_\varepsilon(X) = \{x \in \mathbb{R}^n \mid \|x - X\| < \varepsilon\} \tag{4.6.20}$$

Proof. First notice that by the continuity of f and the compactness of Δ_{n-1} necessarily

$$\operatorname{dist}(x_k, X^*) \to 0 \tag{4.6.21}$$

otherwise for an $\varepsilon > 0$ we could pick a converging subsequence $\{y_k\}$ of $\{x_k\}$ outside $B_\varepsilon(X^*)$ so that $f(y_k) \to f(\lim_{k\to\infty} y_k) > f^*$, a contradiction.
Since the family of sets $\{A_i\}_{1\leq i\leq C}$ is formed by compact and disjoint sets we have that $D = \min_{1\leq i<j\leq C} \operatorname{dist}(A_i, A_j)/2 > 0$. Consider any $\delta < \frac{D}{2}$: then for every $x \in A_i,\ y \in A_j$ we have $\|x - y\| > 2\delta$ so that $\{B_\delta(A_i)\}_{1\leq i\leq C}$ is a family of open and disjoint sets.
Let

$$\tilde{f} = \min\{f(x)\ |x \in \Delta_{n-1}/\bigcup_{i=1}^{C} B_\delta(A_i)\} = \min\{f(x)\ |x \in \Delta_{n-1}/B_\delta(X^*)\} > f^*$$

Since $f(x_k) \to f^*$ there exists $\bar{k}$ such that $f(x_k) < \tilde{f}$ for every $k \geq \bar{k}$. This implies that $x_k \in B_\delta(X^*)$ for every $k \geq \bar{k}$. Let $x_{\bar{k}} \in A_i$. We claim that $x_k \in B_\delta(A_i)$ for every $k \geq \bar{k}$. It suffices to show that if $x_k \in B_\delta(A_i)$ then also $x_{k+1} \in B_\delta(A_i)$. Assume by contradiction $x_{k+1} \in A_j$ with $j \neq i$. Then there exists $p \in \operatorname{conv}(x_k, x_{k+1})$ such that $p \notin \bigcup_{i=1}^{C} B_\delta(A_i)$ because otherwise $\{B_\delta(A_i) \cap \operatorname{conv}(x_k, x_{k+1}\}_{1\leq i\leq C}$ would be a partition open in $\operatorname{conv}(x_k, x_{k+1})$ of a connected segment. But then $f(p) \geq \tilde{f} > f(x_k)$, contradicting hypothesis (4.6.19).
We now have $x_k \in B_\delta(A_i)$ which implies $x_k \notin B_\delta(A_j)$ for every $k \geq \bar{k},\ j \neq i$. This means that $\operatorname{dist}(A_i, x_k) < \delta$ and $\operatorname{dist}(A_j, x_k) > \delta$ for every $j \neq i,\ k \geq \bar{k}$, which implies $\operatorname{dist}(X^*, x_k) = \operatorname{dist}(A_i, x_k)$ for every $k \geq \bar{k}$. To finish, notice that since $\operatorname{dist}(X^*, x_k) \to 0$ then also $\operatorname{dist}(A_i, x_k) \to 0$. □

If ∇f has the Lipschitz property and the step size respect a certain upper bound depending on the Lipschitz constant and the current center x_k then condition (4.6.19) still holds. The proof uses the standard descent lemma (see [6], proposition 6.1.2).

Lemma 4.6.5. *Consider a sequence $\{x_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ such that $x_{k+1} = x_k + \alpha_k d_k$ with $\alpha_k \in \mathbb{R}$, $d_k \in \mathbb{R}^n$. Assume that $0 \leq \alpha_k \leq (-2\nabla f(x_k), d_k)/\|d_k\|^2 L$. Then the sequence $\{x_k\}_{k\in\mathbb{N}}$ has the property* (4.6.19).

Proof. By the standard descent lemma

$$f(x) \leq f(x_k) + (\nabla f(x_k), d) + \frac{L\|d\|^2}{2}$$

so that

$$f(x_k + \alpha d_k) \leq f(x_k) + \alpha(\nabla f(x_k), d_k) + \alpha^2 \frac{L\|d_k\|^2}{2} \tag{4.6.22}$$

Since for $0 \leq \alpha \leq \frac{-2(\nabla f(x_k), d_k)}{L\|d_k\|^2}$ we have

$$\alpha(\nabla f(x_k), d_k) + \alpha^2 \frac{L\|d_k\|^2}{2} \leq 0 \tag{4.6.23}$$

for every $x \in \operatorname{conv}(x_k, x_{k+1}) \subseteq \{x + \alpha_k d_k \mid 0 \leq \alpha \leq \frac{-2(\nabla f(x_k), d_k)}{L\|d_k\|^2}\}$

$$f(x) = f(x + \alpha d_k) \leq f(x_k) + \alpha(\nabla f(x_k), d_k) + \alpha^2 \frac{L\|d_k\|^2}{2} \leq f(x_k) \tag{4.6.24}$$

□

4.6.4 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 simplex. We've already used this property of the AFW in section 3.3 of chapter 2. There we proved that the sequence generated by the AFW is definitely on the face of the polytope exposed by $-\nabla f(x^*)$, where x^* is a point in a certain subset of minimizers to which the function converges. An analogous result was already well known for the gradient projection algorithm, and in section 3.2 of chapter 2 we gave explicit estimates for the convergence rate. In this section our aim is to give explicit estimates for the AFW active set identification complexity for generic polytopes. With respect to the considerations made in section 3.3 we are then more interested in obtaining quantitative results. There is however some overlap between the beginning of this section and section 3.3, which we maintain instead of referencing equations for clarity.
Before stating the general theorem we need to introduce formal notations and prove a few simple properties in the generic polytope setting.
Let $P = \{x \in \mathbb{R}^n \mid Cx \leq b\}$ be a polytope and $f : P \to \mathbb{R}^n$ be a function with gradient having Lipschitz constant L. In the rest of this section the vectors in $\mathbb{R}^n$ have dimension $n \times 1$ and $\nabla f(x)$ has dimension $1 \times n$, so that we can use the product between matrices and omit the scalar product notation.
To define the AFW algorithm we need a finite set of atoms $\mathcal{A}$ such that $\mathrm{conv}(\mathcal{A}) = P$. As for the simplex we can then define for every $a \in \mathcal{A}$ the multiplier function $\lambda_a : P \to \mathbb{R}$ by

$$\lambda_a(x) = \nabla f(x)(a - x)$$

Let finally A be a matrix having for columns the atoms in $\mathcal{A}$, so that A is also a linear transformation mapping $\Delta_{|\mathcal{A}|-1}$ in P with $Ae_i = A^i \in \mathcal{A}$.
In order to apply Theorem 4.2.1 we need to check that the transformed problem

$$\min\{f(Ax) \mid x \in \Delta_{|\mathcal{A}|-1}\} \tag{4.6.25}$$

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:

$$\nabla\tilde{f}(x) - \nabla\tilde{f}(y) = (\nabla f(Ax) - \nabla f(Ay))A \leq L\|A(x-y)\|\|A^T\| \leq L\|A\|\|A^T\|\|x-y\| \tag{4.6.26}$$

This computation also shows that $\nabla\tilde{f}$ has Lipschitz constant

$$L_A = L\|A\|\|A^T\| \tag{4.6.27}$$

Also λ is invariant under affine transformation, meaning that $\lambda_{A^i}(Ax) = \lambda_i(x)$ for every $i \in \{1, ..., |\mathcal{A}|\}$, $x \in P$. Indeed

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

We now need to check that the active set identification property and the disjoint compact partition property used in Theorem 4.3.5 are also invariant under affine transformation. Let X^*_P be the set of minimizers for f on P, so that $X^* = A^{-1}(X^*_P)$ is the set of minimizers for $\tilde{f}$. If $X^*_P = \bigcup_{i=1}^{C} B_i$ with $\{B_i\}_{1\leq i\leq C}$ compact and disjoint than

also $X^* = \bigcup_{i=1}^{C} A^{-1}(B_i)$ with $\{A^{-1}(B_i)\}_{1\leq i\leq C}$ compact and disjoint. 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 $x^* \in X^*_P$ and let

$$P^*(x^*) = \{x \in P \mid \nabla f(x^*)x = \nabla f(x^*)x^*\} = \text{argmax}\{-\nabla f(x^*)x \mid x \in P\} \quad (4.6.28)$$

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

$$P^*(x^*) = \text{conv}\{a \in \mathcal{A} \mid \lambda_a(x^*) = 0\} \quad (4.6.29)$$

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

$$P^*(x^*) = \text{conv}\{A_i \in \mathcal{A} \mid i \in I(x^*)\} \quad (4.6.30)$$

A consequence of (4.6.30) is that given any subset B of P with the active set identification property necessarily $P(x^*) = P(y^*)$ for every $x^*, y^* \in P$, since $I(x^*) = I(y^*)$. For such a subset B we can then define

$$P^*(B) = P^*(x^*) \text{ for any } x^* \in B \quad (4.6.31)$$

where the definition does not depend on the specific $x^* \in B$ considered. We can now restate Theorem 4.3.5 in slightly different terms:

Theorem 4.6.6. *Assume that X^*_P and $\{x_k\}$ have the properties described in 4.3.5. Then there exists M and $i \in \{1, ..., C\}$ such that $x_k \in P^*(A_i)$ for every $k \geq M$.*

Proof. Follows from 4.3.5 and the affine invariance properties discussed above. □

We now generalize the analysis of the strongly convex case.
The technical problem here is that strong convexity, which is used in Corollary 4.3.6, is not maintained by affine transformations, so that instead we will have to use a weaker error bound condition. As a possible alternative, in [34] linear convergence of the AFW is proved with dependence only on affine invariant parameters, so that any version of Theorem 4.2.1 and Corollary 4.3.6 depending on those parameters instead of u, L would not need this additional analysis.
Let x^* be the unique minimum of f on P and $u > 0$ be such that

$$f(x) \geq \frac{u}{2}\|x - x^*\|^2 \quad (4.6.32)$$

The function $\tilde{f}$ inherits the error bound condition necessary for Corollary 4.3.6 from the strong convexity of f: for every $x \in \Delta_{|\mathcal{A}|-1}$ by [3], lemma 2.2 we have

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

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

$$\tilde{f}(x) - \tilde{f}^* = f(Ax) - f(x^*) \geq \frac{u}{2}\|Ax - x^*\|^2 \geq \frac{u}{2\theta^2}\text{dist}(x, X^*)^2 \tag{4.6.33}$$

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

$$\tilde{f}(x) - \tilde{f}^* \geq \frac{u}{2n\theta^2}\text{dist}_1(x, X^*)^2 \tag{4.6.34}$$

where dist_1 is the set point distance computed with respect to $\| \cdot \|_1$.
Having proved this error bound condition for $\tilde{f}$ we can now generalize (4.2.6):

Corollary 4.6.7. *The sequence $\{x_k\}$ generated by the AFW is in $P^*(x^*)$ for*

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

where $f(x_k) - f(x^*) \leq q^k(f(x_0) - f(x^*))$, $u_P = \frac{u}{2n\theta^2}$, $r_* = \frac{\delta_{\min}}{2L+\delta_{\min}}$ with $\delta_{\min} = \min\{\lambda_a(x^*) \mid \lambda_a(x^*) > 0\}$.

Proof. Let $I = \{i \in \{1, ..., |\mathcal{A}|\} \mid \lambda_{A^i}(x^*) = 0\}$, $P^* = P^*(x^*)$. Since $P^* = \text{conv}(\mathcal{A} \cap P^*)$ and by (4.6.30) $\text{conv}(\mathcal{A} \cap P^*) = \text{conv}\{A^i \mid i \in I\}$ the theorem is equivalent to prove that for every k greater than the bound $x_k \in \text{conv}\{A^i \mid i \in I\}$. So if $\{\tilde{x}_k\}$ is the sequence corresponding to $\{x_k\}$ generated by the AFW on the simplex we need to prove that for every k greater than the bound

$$\tilde{x}_k \in \text{conv } \{e_i \mid i \in I\}$$

or in other words $(\tilde{x}_k)_i = 0$ for every $i \in I^c$.
Reasoning as in Corollary 4.3.6 we get that $\text{dist}_1(\tilde{x}_k, X^*) < r_*$ for every

$$k \geq \frac{\ln(h_0) - \ln(u_P r_*^2/2)}{\ln(1/\text{q})}) \tag{4.6.35}$$

Let $\bar{k}$ be the minimum index such that (4.6.35) holds. For every $k \geq \bar{k}$ there exists $y^* \in X^*$ with $\|x_k - y^*\|_1 < r_*$. But $\lambda_i(x) = \lambda_{A^i}(x^*)$ for every $x \in X^*$ by the invariance of λ, so that we can apply Theorem 4.2.1 with fixed point y^* and obtain that if $J_k = \{i \in I^c \mid \tilde{x}_i > 0\}$ then $J_{k+1} \leq \max(0, J_k - 1)$. The conclusion follows exactly as in Corollary 4.3.6.

□

Chapter 5

Appendix

We give in this chapter the necessary definitions and basic theorems used in the rest of the thesis. The author was not able to find a reference for some elementary properties, and in this case the (straightforward) proofs are included.

5.1 Preliminaries

Definition 5.1.1. Given a convex and closed cone $C \subseteq \mathbb{R}^n$ we define:

1. C^d as the dual of C: $C^d = \{x \in \mathbb{R}^n \mid (x, c) \leq 0 \ \forall c \in C\}$
2. $C_\delta = \{0\} \cup \{x \in \mathbb{R}^n \setminus \{0\} \mid \text{dist}(\frac{x}{\|x\|}, C) \leq \delta\}$ for $\delta \geq 0$
3. $C_{-\delta} = \{0\} \cup \{x \in \mathbb{R}^n \setminus \{0\} \mid \text{dist}(\frac{x}{\|x\|}, C^c) \geq \delta\}$ for $\delta \geq 0$

We will need the following results that relate the distance of a point from the dual cone to the norm of the projection on the cone:

Proposition 5.1.2. *For every $x \in \mathbb{R}^n$*

$$\text{dist}(x, C^d) = \sup_{c \in C} (\hat{c}, x) \tag{5.1.1}$$

As stated in [11] this is an immediate consequence of the Moreau decomposition:

$$x = \pi(C, x) + \pi(C^d, x)$$

However here we prove this statement in a way that generalizes straightforwardly to Banach spaces, which will be useful for the analysis in section 5.6.

Proof. First we show that for every $\varepsilon > 0$, $c \in C \setminus \{0\}$

$$\text{dist}(x, C^d) + \varepsilon \geq (\hat{c}, x) \tag{5.1.2}$$

Let $c^* \in C^d$ such that $\text{dist}(c^*, x) \leq \text{dist}(x, C^d) + \varepsilon$. Then

$$(\hat{c}, x) = (\hat{c}, c^* + (x - c^*)) \leq (\hat{c}, c^* - x) \leq \|c^* - x\| \leq \text{dist}(x, C^d) + \varepsilon \tag{5.1.3}$$

where in the first inequality we used $(c^*, c) \leq 0$. This proves (5.1.2).
It remains to prove that for every $\varepsilon > 0$ or equivalently for every $\varepsilon \in (0, \mathrm{dist}(x, C^d)/2)$ there exists $c \in C \setminus \{0\}$ such that

$$\mathrm{dist}(x, C^d) - \varepsilon \leq (\hat{c}, x) \tag{5.1.4}$$

Let $d = \mathrm{dist}(x, C^d) - \varepsilon/2$. Consider c separating the open convex set $B(x, d + \varepsilon/4)$ and C^d, so that

$$c \in C \setminus \{0\}, \ (c, b) \geq 0 \tag{5.1.5}$$

for every $b \in B(x, d + \varepsilon/4)$. Then

$$(\hat{c}, x) = (\hat{c}, x - d\hat{c} + d\hat{c}) = (\hat{c}, x - d\hat{c}) + d \geq d > \mathrm{dist}(x, C^d) - \varepsilon \tag{5.1.6}$$

where in the first inequality we used that $x - d\hat{c} \in B(x, d + \varepsilon/4)$. Hence (5.1.4) is proved so that combining it with (5.1.2) and taking the limit for $\varepsilon \to 0$ we get (5.1.1). □

Remark 5.1.3. The sup in (5.1.1) is actually a max for cones in $\mathbb{R}^n$ by compactness.

It is not difficult to prove that C_δ and $C_{-\delta}$ are cones for every $\delta > 0$, and moreover that $C_{-\delta}$ is the closed convex cone dual of C_δ^d. In particular, in chapter 1 we use the following:

Proposition 5.1.4. *Given a cone C:*
a) C_δ is a cone for every $\delta \in [-1, 1]$.
b) If C is closed and convex, $(C_\delta^d)^d \supseteq C_{-\delta}$ for every $1 \geq \delta > 0$.

Proof. a) We assume $\delta \geq 0$, but the same proof works also for $\delta < 0$. By definition, $\{0\} \in C_\delta$. If $x \neq 0 \in C_\delta$ then for every $\lambda > 0$ we have

$$\mathrm{dist}(\frac{\lambda x}{\|\lambda x\|}, C) = \mathrm{dist}(\frac{x}{\|x\|}, C) \leq \delta \tag{5.1.7}$$

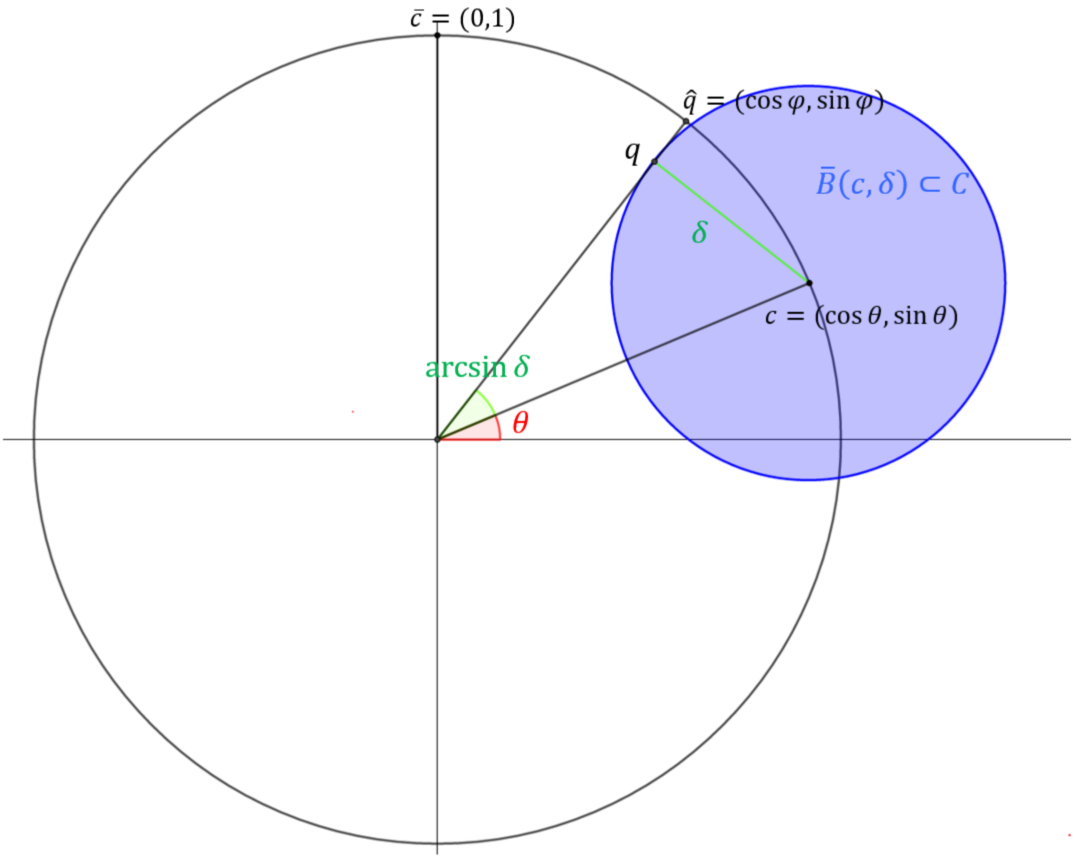
so that $\lambda x \in C$.
b) Proving the inclusion is equivalent to prove that given $c \in C_{-\delta}$ for every $\bar{c} \in C_\delta^d$ we have $(\bar{c}, c) \leq 0$. Fix $c \in C_{-\delta}$ and $\bar{c} \in C_\delta^d$. Without loss of generality we can assume $\|c\| = \|\bar{c}\| = 1$. By Proposition 5.1.2 we have

$$\mathrm{dist}(\bar{c}, C^d) \leq \delta \Rightarrow \max_{\hat{v} \in C} (\hat{v}, \bar{c}) \leq \delta \tag{5.1.8}$$

Assume by contradiction that $(c, \bar{c}) > 0$. We identify the plane containing $c, \bar{c}, 0$ with $\mathbb{R}^2$, and assume without loss of generality that $\bar{c} = (0, 1)$, $c = (\cos(\theta), \mathrm{sen}(\theta))$ for some $0 < \theta \leq \frac{\pi}{2}$. We distinguish two cases.
Case 1: $\bar{B}(c, \delta) \cap \{(0, y) \mid y > 0\} \neq \emptyset$. Then since $\bar{B}(c, \delta) \subset C$ we also have $\bar{c} = (0, 1) \in C$. Then the max in (5.1.8) is 1 attained for $\hat{v} = \bar{c}$. It follows $\delta = 1$ so that $C_\delta^d = C_1^d = \mathbb{R}^n$, contradiction because then $C = C_{-1} = C_{-\delta} = \{0\}$ so that $c \notin C_{-\delta}$.
Case 2: $\bar{B}(c, \delta) \cap \{(0, y) \mid y > 0\} = \emptyset$. We refer to Figure 5.1 for the analysis of this case. We have for $\varphi = \arcsin^{-1}(\delta) + \theta \in (0, \frac{\pi}{2})$

$$\hat{q} \stackrel{\mathrm{def}}{=} (\cos(\varphi), \sin(\varphi)) \in C \tag{5.1.9}$$

Figure 5.1: Configuration of the proof that $C_{-\delta} \subset (C_\delta^d)^d$.



using again that $\bar{B}(c,\delta) \subset C$. But then

$$(\bar{c}, \hat{q}) = \sin(\varphi) = \sin(\arcsin^{-1}(\delta) + \theta) > \sin(\arcsin^{-1}(\delta)) = \delta$$

contradicting (5.1.8).

□

Studying convex problems over a polytope with a degenerate solution requires relating conic combinations to the facial structure of the cone, and also a certain lemma about the relation between the faces of a cone C and the faces of his dual C^d.

Proposition 5.1.5. *Let $M(C)$ be a $n \times m$ matrix whose columns generate the convex cone C. Then for every face F of C if $M_F = \{i \in \{1, ..., m\} \mid M(C)_i \in F\}$ in F the relative interior of F is the set of vectors*

$$\{c = \sum_{i \in M_F} \lambda_i c_i \mid \lambda_i > 0 \forall i \in M_F\}$$

Proof. See for instance [12]. □

Theorem 5.1.6. *If C is a convex polyhedral cone then there is a one to one idempotent and inclusion reversing correspondence*

$$\{\text{faces of } C\} \longleftrightarrow \{\text{faces of } C^d\}$$
$$\tau \longleftrightarrow \tau^* = C^d \cap \tau^\perp$$

Proof. See [16], Theorem 9. □

We recall the definition of tangent cone to a general set Ω:

Definition 5.1.7. A vector $w \in \mathbb{R}^n$ is tangent to a set Ω in $\bar{x} \in \Omega$, written $w \in T_\Omega(\bar{x})$, if there exists a sequence of positive scalars $\tau_k \to 0$ and a sequence $\{x_k\}_{k\in\mathbb{N}}$ in Ω with $x_k \to \bar{x}$ such that

$$(x_k - \bar{x})/\tau_k \to w$$

The tangent cone $T_\Omega(\bar{x})$ to Ω in $\bar{x}$ defined as the set of vectors tangent to Ω in $\bar{x}$.

We also recall the general definition of regular normal cone:

Definition 5.1.8. The regular normal cone $\hat{N}(\bar{x})$ to a set Ω in $\bar{x}$ is defined as the set of vectors v satisfying

$$(v, x - \bar{x}) = o(\|x - \bar{x}\|) \text{ for } x \to \bar{x} \text{ in } \Omega \tag{5.1.10}$$

The next proposition characterizes the tangent cone and the normal cone for convex closed sets.

Proposition 5.1.9. *Let Ω be a closed convex set. For every point $\bar{x} \in \Omega$ if $T_\Omega(\bar{x})$ is the tangent cone to Ω in $\bar{x}$ then*

$$\begin{aligned} T_\Omega(\bar{x}) &= \mathrm{cl}\{w \mid \exists\lambda > 0 \text{ with } x + \lambda w \in \Omega\}, \\ \mathrm{int}(T_\Omega(\bar{x})) &= \{w \mid \exists\lambda > 0 \text{ with } x + \lambda w \in \ \mathrm{int}(\Omega)\} \\ \hat{N}_\Omega(\bar{x}) &= T_\Omega(\bar{x})^d \end{aligned}$$

Proof. See [40], Theorem 6.9. □

Using these characterizations we now prove a formula connecting the maximal "slope" of a linear function along an admissible direction to the tangent and the normal cone:

Proposition 5.1.10. *If Ω is a closed convex subset of $\mathbb{R}^n$, $\bar{x} \in \Omega$ then for every $r \in \mathbb{R}^n$*

$$\max\{0, \sup_{h\in\Omega\setminus\{\bar{x}\}}(r, \frac{h-\bar{x}}{\|h-\bar{x}\|})\} = \mathrm{dist}(r, \hat{N}_\Omega(\bar{x})) = \|\pi(T_\Omega(\bar{x}), r)\|$$

Proof. We first prove that

$$\sup_{h\in\Omega/\{\bar{x}\}}(r, \frac{h-\bar{x}}{\|h-\bar{x}\|}) = \sup_{h\in T_\Omega(\bar{x})/\{0\}}(r, \hat{h}) \tag{5.1.11}$$

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

$$\mathrm{cl}(\{\frac{h-x}{\|h-x\|} \mid h \in \Omega \setminus \{0\}\}) = \{\hat{h} \mid h \in T_\Omega(\bar{x})/\{0\}\} \tag{5.1.12}$$

and (5.1.11) follows immediately by the continuity of $(r, \cdot)$.
Since $\hat{N}_\Omega(\bar{x}) = T_\Omega(\bar{x})^d$ the first equality is exactly the one of Lemma 5.1.2 if $r \notin \hat{N}_\Omega(\bar{x})$,

and it is trivial since both terms are clearly 0 if $r \in \hat{N}_\Omega(x)$.
It remains to prove

$$\max\{0, \sup_{h \in T_\Omega(\bar{x})/\{0\}} (r, \hat{h})\} = \|\pi(T_\Omega(\bar{x}), r)\| \tag{5.1.13}$$

Let π_r be the projection of r on $T_\Omega(\bar{x})$, and π_r^d the projection of r on $\hat{N}_\Omega(\bar{x})$. By the Moreau - Yosida decomposition $r = \pi_r + \pi_r^d$ with $\pi_r \perp \pi_r^d$. If $r \in \hat{N}_\Omega(\bar{x})$ then $\pi_r = 0$ so that again the equation (5.1.13) is true with both sides equal to 0.
Otherwise on the one hand

$$\sup_{h \in T_\Omega(\bar{x})/\{0\}} (r, \hat{h}) \geq (r, \hat{\pi}_r) = (\pi_r + \pi_r^d, \hat{\pi}_r) = \|\pi_r\| \tag{5.1.14}$$

and on the other hand for every $h \in T_\Omega(\bar{x}) \setminus \{0\}$

$$(\hat{h}, r) = (\hat{h}, \pi_r + \pi_n) \leq (\hat{h}, \pi_r) \leq \|\pi_r\| \tag{5.1.15}$$

Taking the sup in (5.1.15) and combining it with (5.1.14) we get the desired equality.
□

5.2 Length function and vertex width diameter ratio.

The width of a convex set, and more precisely its ratio with the diameter, will be crucial to give lower bounds for the decrease of the objective function at every step in the convergence analysis of the FDFW.

Definition 5.2.1. Given $\Omega \subset \mathbb{R}^n$ closed and convex we define the directional width of Ω in y with respect to $r \in \mathbb{R}^n/\{0\}$ as

$$\mathrm{dirW}(\Omega, r, x) = \max\{r(x - y) \mid x \in \Omega\}$$

the directional width of Ω with respect to r as

$$\mathrm{DirW}(\Omega, r) = \max\{r(x - y) \mid x, y \in \Omega\}$$

and the width of Ω as

$$W(\Omega) = \min\{\mathrm{DirW}(\Omega, r) \mid r \in \partial B_{\mathbb{R}^n}(0, 1)\}$$

It is easy to check that this width function is non negative and monotone increasing with respect to the inclusion. The width of a set isn't of much help in the convergence analysis of the FDFW method when dealing with points on the boundary, so that for these we need to define a sort of width and diameter dependent on a restricted set of admissible directions. We define these parameters in section 2.4.4. The motivation behind definitions 2.4.7, 2.4.8 is fundamentally that they allow us to give a lower bound on the directional normalized width for every vertex and every direction.

We now prove a few properties of the parameters defined in section 2.4.4. Let $p_{\Omega,\bar{x}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be the Minkowski functional of $\Omega_{\bar{x}} = \Omega - \{\bar{x}\}$:

$$p_{\Omega,\bar{x}}(x) = \inf\{\alpha \in \mathbb{R}_{>0} \cup \{\infty\} \mid \alpha^{-1}x \in \Omega - \{\bar{x}\}\}$$

It is well known that this functional is convex, homogeneous for positive scalars and lower semicontinuos on all $\mathbb{R}^n$. It is now convenient to use the convention $1/0 = \infty$, $1/\infty = 0$ in the rest of this section. With this convention it easy to prove the following:

$$l_{\Omega,\bar{x}}(c) = \frac{1}{p_{\Omega,\bar{x}}(x)} \quad \forall x \in T_C(\bar{x}) \tag{5.2.1}$$

since for a fixed direction $c \in T_C(\bar{x})/\{0\}$ one has

$$l_{\Omega,\bar{x}}(c) = l_{\Omega,\bar{x}}(\hat{c})/\|c\| = 1/p_{\Omega,\bar{x}}(x) \tag{5.2.2}$$

As a corollary, to the properties of $p_{\Omega,\bar{x}}$ correspond analogous properties for $l_{\Omega,\bar{x}}$, as we prove in the following proposition.

Proposition 5.2.2. *With the notation introduced above, let $C = T_\Omega(\bar{x})$:*

1. *$l_{\Omega,\bar{x}}$ is continuous in* $\mathrm{int}(C)$*, upper semicontinuous in C;*
2. *For $0 < \beta$ the* inf *is actually a* min *in the definition* (2.4.35) *of $l^b_{\Omega,\bar{x}}$; the* sup *is always a* max *in the definition* (2.4.35) *of $l^B_{\Omega,\bar{x}}$;*
3. *The following formula holds for Ω strictly convex:*

$$R_{\Omega,\bar{x}}(k) = \inf_{0<\delta\leq 1} \frac{\min\{p_{\Omega,\bar{x}}(\hat{c}) \mid 0 \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}}{\max\{p_{\Omega,\bar{x}}(\hat{c}) \mid k\delta \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}}$$

Proof. 1) The continuity of $l_{\Omega,\bar{x}}$ in $\mathrm{int}(C)$ follows immediately if we can prove that $p_{\Omega,\bar{x}}$ is continuous in $\mathrm{int}(C)$. By Proposition 5.1.9 for every $x \in C \setminus \{0\}$ there exists $\lambda > 0$ such that $\lambda x \in \Omega - \{\bar{x}\}$. Then

$$p_{\Omega,\bar{x}}(x) = \inf\{\alpha \in \mathbb{R}_{\geq 0} \cup \{\infty\} \mid \alpha^{-1}x \in \Omega - \{\bar{x}\}\} \leq \frac{\lambda}{\|x\|} < +\infty \tag{5.2.3}$$

so that $p_{\Omega,\bar{x}}$ is finite in $T_\Omega(\bar{x})$. But $p_{\Omega,\bar{x}}$ is also convex, hence continuous in its domain which contains C°.
As for the upper semicontinuity, it follows from the fact that the inverse of a positive lower semicontinuous function is an upper semicontinuous function.
2) Follows immediately from the fact that an upper semicountinuous function and a lower semicontinuous function have always a maximum and a minimum respectively on a compact set.
3) We have

$$\begin{aligned} R_{\Omega,\bar{x}}(k) &= \inf_{0<\delta\leq 1} \frac{\inf\{l_{\Omega,\bar{x}}(\hat{c}) \mid k\delta \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}}{\sup\{l_{\Omega,\bar{x}}(\hat{c}) \mid 0 \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}} = \inf_{0<\delta\leq 1} \frac{\inf\{1/p_{\Omega,\bar{x}}(\hat{c}) \mid k\delta \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}}{\sup\{1/p_{\Omega,\bar{x}}(\hat{c}) \mid 0 \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}} = \\ &= \inf_{0<\delta\leq 1} \frac{\inf\{p_{\Omega,\bar{x}}(\hat{c}) \mid 0 \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}}{\sup\{p_{\Omega,\bar{x}}(\hat{c}) \mid k\delta \leq \mathrm{dist}(\hat{c}, C^c) \leq \delta\}} \end{aligned} \tag{5.2.4}$$

where in the last term the inf is actually a min by lower semicontinuity of $p_{\Omega,\bar{x}}$ and the sup is actually a max by the continuity of $p_{\Omega,\bar{x}}$ on C° proved in point 1). $\square$

The following property concerning unitary vectors maximizing the distance from the boundary of a cone will be useful to study NW(Ω) for strictly convex sets.

Proposition 5.2.3. *If C is a full dimensional closed convex cone different from $\mathbb{R}^n$, then*

$$\operatorname{argmin}_{c\in\mathbb{R}^n\setminus\{0\}}\operatorname{dist}(\hat{c}, C^c) \subset C^\circ \cap -C^d \tag{5.2.5}$$

The proof is a fairly straightforward application of optimality conditions. However, transforming the problem in a constrained programming one to which we can apply optimality conditions requires a few observations.

Proof. As first step we want to find a problem equivalent to minimizing the distance from the boundary to which we can apply Kuhn Tucker necessary optimality conditions (see [2], Theorem 3.7).
We begin by rewriting the problem equivalently as

$$\sup\{\operatorname{dist}(c, C^c) \mid c \in \partial B(0,1)\} = \max\{\operatorname{dist}(c, C^c) \mid c \in \partial B(0,1)\} \tag{5.2.6}$$

where the sup is actually a max because we are maximizing a continuous function in a compact set. We also have by the positive homogeneity of the distance

$$\lambda \max\{\operatorname{dist}(c, C^c) \mid c \in \partial B(0,1)\} = \max\{\operatorname{dist}(c, C^c) \mid c \in \lambda\partial B(0,1)\} \tag{5.2.7}$$

so that

$$\begin{aligned}\max\{c \in \partial B(0,1) \mid \operatorname{dist}(c, C^c)\} &= \max\{\operatorname{dist}(c, C^c) \mid c \in \lambda\partial B(0,1), 0 \le \lambda \le 1\} \\ &= \max\{\operatorname{dist}(c, C^c) \mid c \in \bar{B}(0,1)\}\end{aligned} \tag{5.2.8}$$

with $\operatorname{argmax}\{\operatorname{dist}(c, C^c) \mid c \in \bar{B}(0,1)\} \subset \partial B(0,1)$. Indeed if there existed $x \in \mathring{B}(0,1) \setminus \{0\} \in \operatorname{argmax}\{\operatorname{dist}(c, C^c) \mid c \in \bar{B}(0,1)\}$ we would have $\operatorname{dist}(\hat{x}, C^c) = \frac{1}{\|x\|}\operatorname{dist}(x, C^c) > \operatorname{dist}(x, C^c)$. We also have clearly

$$\operatorname{argmax}\{\operatorname{dist}(c, C^c) \mid c \in B(\bar{0,1})\} \subset C^\circ$$

so that in particular $0 \notin \operatorname{argmax}\{\operatorname{dist}(c, C^c) \mid c \in \bar{B}(0,1)\}$ since by hypothesis $0 \notin \mathring{C}$.
By the biduality theorem for cones

$$C = (C^d)^d = \bigcap_{c\in C^d\setminus\{0\}} \{x \in \mathbb{R}^n \mid (\hat{c}, x) \le 0\} = \bigcap_{c\in C^d\cap\partial B(0,1)} A_c$$

where $A_c = \{x \in \mathbb{R}^n \mid (c, x) \le 0\}$. Therefore

$$\operatorname{dist}(x, C^c) = \operatorname{dist}(x, (\bigcap_{c\in C^d\cap\partial B(0,1)} A_c)^c) = \operatorname{dist}(x, (\bigcup_{c\in C^d\cap\partial B(0,1)} A_c^c)) = \inf_{c\in C^d\cap\partial B(0,1)} \operatorname{dist}(x, A_c^c) \tag{5.2.9}$$

Let $x \in C \cap \bar{B}(0,1)$ and $c \in \partial B(0,1) \cap C^d$. We have

$$\operatorname{dist}(x, A_c^c) = \|\pi(r_c, x)\| \tag{5.2.10}$$

where r_c is the line generated by c. Since $c \in C^d$ we have $(-c, x) \ge 0$ which together with $\|c\| = 1$ implies

$$\|\pi(r_c), x\| = (-c, x) \tag{5.2.11}$$

Starting from (5.2.9) we can now further transform the expression for $\mathrm{dist}(x, C^c)$:

$$\inf_{c \in C^d \cap \partial B(0,1)} \mathrm{dist}(x, A_c^c) = \inf_{c \in C^d \cap \partial B(0,1)} -(c, x) = \min_{c \in C^d \cap \partial B(0,1)} -(c, x) \tag{5.2.12}$$

where the last equality is justified by the continuity of $(x, \cdot)$ and the compactness of

$$C^d \cap \partial B(0,1)$$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f(x) = \max_{c \in C^d \cap \partial B(0,1)} (c, x)$$

Since $f(x)$ is the maximum of convex continuous functions it is convex, l.s.c and

$$\partial f(x) = \mathrm{conv}\{c \in C^d \cap \partial B(0,1) \mid (c, x) = f(x)\} \subset C^d \tag{5.2.13}$$

as it follows from the formula for the subdifferential of the max of convex functions. Moreover, for every $x \in C^c$ by the separation theorems there exists $c \in C^d$ such that $(c, x) \geq 0$, so that $f(x) \geq 0$. At the same time by (5.2.9) and (5.2.8) we have $f(x) = -\mathrm{dist}(x, C^c)$ for every $x \in C$. Then

$$\mathrm{argmin}_{x \in \bar{B}(0,1)} f(x) = \mathrm{argmax}_{x \in \bar{B}(0,1)} \mathrm{dist}(x, C^c) \subset C^\circ$$

and we can finally rewrite our problem as

$$\max\{\mathrm{dist}(c, C^c) \mid c \in \bar{B}(0,1)\} = \max\{-f(c) \mid c \in \bar{B}(0,1)\} = -\min\{f(c) \mid \|c\|^2 \leq 1\} \tag{5.2.14}$$

As observed at the beginning of the proof if $\bar{c} \in C \cap \bar{B}(0,1)$ is a solution necessarily $\|\bar{c}\| = 1$, so that the Kuhn Tucker optimality conditions dictates

$$0 \in \partial f(\bar{c}) + \lambda 2\bar{c} \tag{5.2.15}$$

for a certain $\lambda \geq 0$. Since 0 is a vertex of the cone C^d it can not be a proper conic combination of elements in C^d different from 0, hence by (5.2.13) $0 \notin \partial f(\bar{c})$. But then $\lambda > 0$ in (5.2.15) so that $\bar{c} \in -\partial f(\bar{c})/2\lambda$, which implies the thesis because $-\partial f(\bar{c}) \subset -C^d$. $\square$

5.3 Step size oracle for convex hulls

Let $\Omega = \mathrm{conv}(\mathcal{A})$, with

$$\mathcal{A} = \{a_i \mid 1 \leq i \leq n\} \tag{5.3.1}$$

a finite subset of $\mathbb{R}^n$. We are interested in trasforming the problem

$$\alpha_{\max}(\bar{x}, d) = \max\{\alpha \in \mathbb{R} \mid \bar{x} + \alpha d \in \Omega\} \tag{5.3.2}$$

in a linear programming problem for a fixed $\bar{x} \in \Omega$, $d \in \mathbb{R}^n$.
Consider an hyperplane

$$H_{c,\beta} = \{x \in \mathbb{R}^n \mid (c, x) + \beta = 0\} \tag{5.3.3}$$

such that Ω is contained in the negative half space $H^-_{c,\beta}$:

$$(c, x) + \beta \leq 0 \ \forall x \in \Omega \tag{5.3.4}$$

Assume also that $(c, d) = 1$. We now consider the set H of hyperplanes with these two properties to give an upper bound on $\alpha_{\max}$. Then we will show that this upper bound coincide with the actual value of $\alpha_{\max}$.
Let $\alpha^{c,\beta}_{\max}(\bar{x}, d)$ be the maximal feasible step from $\bar{x}$ in the direction d with respect to the set $H_{c,\beta}$:

$$\alpha^{c,\beta}_{\max}(\bar{x}, d) = \max\{\alpha \in \mathbb{R} \mid \bar{x} + \alpha d \in H_{c,\beta}\} \tag{5.3.5}$$

From now on we will write $\alpha^{c,\beta}_{\max}$ and $\alpha_{\max}$ instead of $\alpha^{c,\beta}_{\max}(\bar{x}, d)$ and $\alpha_{\max}(\bar{x}, d)$ since $\bar{x}$, d are fixed anyway.
Notice that

$$(c, \bar{x} + \alpha^{c,\beta}_{\max} d) + \beta = 0 \Rightarrow \alpha^{c,\beta}_{\max} = -\beta - (c, \bar{x}) \tag{5.3.6}$$

where we used $(c, d) = 1$.
$\alpha^{c,\beta}_{\max}$ gives an upper bound for $\alpha_{\max}$:

$$\alpha^{c,\beta}_{\max} \geq \alpha_{\max} \tag{5.3.7}$$

because by hypothesis $H^-_{c,\beta} \supseteq \Omega$.
We can now define a linear programming problem which has optimal value $\alpha_{\max}$:

$$\begin{aligned} z = \min \ & -(c, \bar{x}) - \beta \\ & (c, d) = 1 \\ & (c, a_i) + \beta \leq 0 \ \forall \ 1 \leq i \leq n \end{aligned} \tag{5.3.8}$$

Since $\Omega = \operatorname{conv}(\{a_i\}_{1 \leq i \leq n})$, the third condition is equivalent to $\Omega \subset H^-_{c,\beta}$. Therefore c, b satisfies the constraint of problem (5.3.8) iff $H_{c,\beta} \in H$. We can now use the above reasoning to conclude that $\alpha_{\max} \leq \alpha^{c,\beta}_{\max} = -(c, \bar{x}) - \beta$ for every feasible (c, β), so that $z \geq \alpha_{\max}$.
To see that equality holds we distinguish two cases. If d is not a feasible direction then $\alpha_{\max} = z = 0$, where the optimal value in problem (5.3.8) is obtained for the hyperplane separating Ω from the ray $\bar{x} + \lambda d$, $\lambda \geq 0$. If d is a feasible direction then $\mathcal{F}(\bar{x} + \alpha_{\max} d) \subsetneq \mathcal{F}(\bar{x})$ so that there exists a supporting plane $H_{\bar{c},\bar{\beta}}$ for Ω in $\bar{x} + \alpha_{\max} d$ not containing $\bar{x}$. But then

$$(\bar{c}, \bar{x} + \alpha_{\max} d) + \beta > (\bar{c}, \bar{x}) + \beta \tag{5.3.9}$$

so that $(\bar{c}, d) > 0$. It is now immediate to check that by diving the coefficients of $H_{\bar{c},\bar{\beta}}$ by $(\bar{c}, d)$ we get a feasible point (c', β') for problem (5.3.8) for which $\alpha_{\max} = -(c', \bar{x}) - \beta'$. This proves that $\alpha_{\max}$ is indeed the optimal value of problem (5.3.8).

5.4 Minimal face for convex hulls

We now show that when $\Omega = \operatorname{conv}(\mathcal{A})$ as in the previous section for every $\bar{x} \in \Omega$ finding the minimal face $\mathcal{F}(\bar{x})$ of Ω containing $\bar{x}$ is equivalent to finding a point in the interior of a certain polyhedron. The key property our analysis is based on is the following: given a polyhedron P and $\bar{x} \in P$ then $\mathcal{F}(\bar{x})$ is the face exposed by the vectors in $\mathrm{ri}(N_P(\bar{x}))$. In formulas, given $r \in \mathrm{ri}(N_P(\bar{x}))$:

$$\mathcal{F}(\bar{x}) = \{y \in P \mid (r, y) = (r, \bar{x})\} \tag{5.4.1}$$

and conversely if $r \in \mathbb{R}^n$ is such that (5.4.1) holds then $r \in \mathrm{ri}(N_P(\bar{x}))$. We use this property extensively also in chapter 32, where we provide references and show the connection with other properties of normal cones.
Applying this property to our problem we get that if we can find $r \in \mathrm{ri}(N_\Omega(\bar{x}))$ then

$$\mathcal{F}(\bar{x}) = \{y \in \Omega \mid (r, \bar{x}) = (r, y)\} = \mathrm{conv}(\{a \in \mathcal{A} \mid (a, r) = (a, \bar{x})\}) \tag{5.4.2}$$

so that we can identify the subset of $\mathcal{A}$ given by its intersection with $\mathcal{F}(\bar{x})$ and consequently solve linear optimization problems checking the value of the objective on the points in this subset.
We have thus reduced the problem of identifying $\mathcal{A} \cap \mathcal{F}(\bar{x})$ to the problem of finding $r \in \mathrm{ri}(N_\Omega(\bar{x}))$. By Proposition 5.1.9

$$N_\Omega(\bar{x}) = \{y \in \mathbb{R}^n \mid (y, x - \bar{x}) \leq 0 \; \forall \; x \in \Omega\} = \{y \in \mathbb{R}^n \mid (y, a) \leq (y, \bar{x}) \; \forall \; a \in \mathcal{A}\} \tag{5.4.3}$$

where we used $\Omega = \mathrm{conv}(\mathcal{A})$ in the second equality. We can finally write inequalities describing $N_\Omega(\bar{x})$ in matrix form

$$N_\Omega(\bar{x}) = \{x \in \mathbb{R}^n \mid \mathcal{A}x \leq e(x, \bar{x})\} \tag{5.4.4}$$

The problem of finding a point in the relative interior of $N_\Omega(\bar{x})$ is then a particular case of the problem of finding a point in the relative interior of a polyhedron (more precisely a cone) given in standard inequality form. This problem has already been studied for its relevance in the initialization of interior point methods and in compressed sensing (see for instance [37], [15] where an algorithm with linear convergence and asymptotic quadratic convergence is given).

5.5 Strictly convex sets

We now prove a lemma guaranteeing a sort of continuity modulus for bounding the distance of an approximated solution from the minimizer of a linear function.

Proposition 5.5.1. *Let Ω be a strictly convex set. Then there exists an increasing function $m_\Omega : [0, +\infty) \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ continuous in 0 such that $m_\Omega(0) = 0$ and*

$$\mathrm{dist}(x, \mathrm{argmin}_{y \in \Omega}(r, y)) \leq m_\Omega\left(\frac{(r, x) - \min_{y \in \Omega}(r, y)}{\|r\|}\right) \tag{5.5.1}$$

for every $x \in \Omega$, $r \in \mathbb{R}^n \setminus \{0\}$.

Proof. Since both sides of equations (5.5.1) are invariant if r is multiplied by a positive scalar, we can assume $r \in \partial B(0, 1)$. We define $c(r)$ as the continuous function from $\partial B(0, 1)$ to $\partial\Omega$ associating to r the minimizer of $(r, \cdot)$ on Ω. Then for $\delta \geq 0$ we define

$$C(r, \delta) = \{y \in \Omega \mid (y - c(r), r) = \delta\} \tag{5.5.2}$$

and

$$m(r,\delta) = \sup\{\|y - c(r)\| \mid y \in C(r,\delta)\} \tag{5.5.3}$$

Thus in particular $m(r,0) = 0$ and the sup is actually a max whenever the set $C(r,\delta)$ is non empty. We finally define

$$m_\Omega(\delta) = \sup_{\substack{r\in\partial B(0,1)\\ 0\le\delta'\le\delta}} m(r,\delta') \tag{5.5.4}$$

for every $\delta > 0$. Since $m(r,0) = 0$ for every $r \in \partial B(0,1)$ we have $m_\Omega(0) = 0$ and m_Ω is also increasing because the sup is taken on increasing sets. It remains to prove that m_Ω is continuous in 0. Assume by contradiction that this is not the case, or equivalently that there exists a sequence $\{r_i\}_{i\in\mathbb{N}}$ in $\partial B(0,1)$ and a sequence $\{\delta_i\}_{i\in\mathbb{N}}$ converging to 0 such that

$$m(r_i,\delta_i) > \varepsilon \tag{5.5.5}$$

for some fixed $\varepsilon > 0$. By compactness we can assume $r_i \to \bar{r}$. Let $s_i \in C(r_i,\delta_i)$ such that

$$\|s_i - c(r_i)\| > \varepsilon \tag{5.5.6}$$

Modulo considering a subsequence we can again assume $s_i \to \bar{s} \in \Omega$. Then on the one hand passing to the limit (5.5.6) we get by the continuity of $c(r)$

$$\|\bar{s} - c(\bar{r})\| > \varepsilon \tag{5.5.7}$$

and on the other hand passing to the limit in $s_i \in C(r_i,\delta_i)$ we get

$$(c(\bar{r}) - \bar{s}, \bar{r}) = 0 \tag{5.5.8}$$

which is incompatible with (5.5.7) since Ω is strictly convex and $c(\bar{r})$ is a point on the boundary. $\square$

5.6 Generalizing on Banach spaces

In this section we discuss a few technical details about how the elements introduced in the preliminaries generalize to Banach spaces. We then use these results to analyze the FDFW method on a more general setting than $\mathbb{R}^n$.

The main references for this section are [2] for convex analysis, and [10] for some elementary properties of Banach spaces.

X will be a Banach space with norm $\|\cdot\|$ and dual X^*. For every $c \in X^*$, $x \in X$ we will write (c,x) instead of $c(x)$.

The notion of (strongly) convex function generalizes in a straightforward way to Banach spaces. As for differentiability, we remark that for a function $f : X \to \mathbb{R}$ we will use the Frechét definition of differential $Df : X \to X^*$. So we say that f is differentiable in a point x if there exists $Df(x) \in X^*$ such that

$$f(x+h) = f(x) + Df(x)(h) + o(\|h\|)$$

We say that f is differentiable in Ω if it is differentiable for every point in a neighborhood of Ω. The Lipschitz condition than becomes $\|Df(x) - Df(y)\| \le L\|x - t\|$

where we are using $\|\cdot\|$ also for the dual norm on X^*. In particular we have that if a function has Lipschitz differential than a fortiori its restriction to any line has Lipschitz derivative. To see this, consider $x \in X$, $d \in X$ with $\|d\| = 1$. Then:

$$\begin{aligned} &\frac{\partial}{\partial t}(f(x+td))|_{t=\bar{t}} - \frac{\partial}{\partial t}f(x+td)|_{t=0} = \\ =&(Df(x+td), d) - (Df(x), d) \leq \|Df(x+td) - Df(x)\|\|d\| \leq Lt \end{aligned} \tag{5.6.1}$$

Another important element of our analysis is the polar of a set, which generalizes the concept of dual cone. Given $A \subset X$, the polar of A is the set $A^d \subset X^*$ satisfying:

$$A^d = \{x^* \in X^* \mid \sup_{x \in A}(x^*, x) \leq 1\} \tag{5.6.2}$$

Before stating its properties we need to recall an important theorem about reflexive Banach spaces:

Proposition 5.6.1. *If a Banach space X is reflexive than every closed, convex and bounded subset of X is weakly compact.*

The properties of the polar set and its connection with conjugate functions are discussed in [2], but here we are interested only in the case where A is closed and convex, with $0 \in \mathrm{int}(A)$. We recall that a closed convex subset of a Banach space is called *smooth* if for every point on the boundary the normal cone is a ray. Here we are using normal cones defined extending the definition of normal cones in $\mathbb{R}^n$ in an obvious way.

Proposition 5.6.2. *Let A be a closed convex subset of a reflexive Banach space such that $0 \in A$. Then*

1. $(A^d)^d = A$
2. $0 \in \mathrm{int}(A) \Leftrightarrow A^d$ *is bounded, and conversely.*
3. *A is (smooth) strictly convex if and only if A^d is (strictly convex) smooth.*

Proof. 1. Obvious corollary of the bipolar theorem. See for instance [2], Theorem 2.26.
2. This is straightforward and well known for $\mathbb{R}^n$ (see for instance exercise B.15, [4]), and in this setting can be proved in the same way.
3. See [2], Theorem 1.101. Even if this theorem is for balls with respect to a certain norm, the proof works step by step also in our setting. □

Finally, if Ω is strictly convex, smooth, and $0 \in \mathrm{int}(\Omega)$ then we can define a generalization of the duality function (which is usually defined for balls) $J : \partial\Omega \to \partial\Omega^d$ imposing $(J(x), x) = 1$, and its inverse $J^* : \partial\Omega \to \partial\Omega^d$ analogously.

Proposition 5.6.3. *Assume that Ω is strictly convex, smooth, bounded and that $0 \in \mathrm{int}(\Omega)$. Then $J(x)$ and $J^*(x)$ are well defined bijections with $J^* = J^{-1}$.*

Proof. It suffices to show that J and J^* are well defined since then $(J(x), x) = 1 \Rightarrow x = J^*(J(x))$ and analogously $y = J(J^*(y))$ which suffices to prove the bijection property. We need to show that for every $x \in \partial\Omega$ there exists a unique $x^* \in \partial\Omega^d$ such that $(x, x^*) = 1$. By the separation theorem there exists x^* such that $(x^*, x) = 1$, $(x^*, y) \leq 1$ for every $y \in \Omega$. Then by definition of Ω^d we have $x^* \in \partial\Omega^d$ and by the smoothness hypotheses necessarily $N_\Omega(x) = \text{cone}(x^*)$ so that x^* is unique. As for J^*, notice that it is the same functional defined for Ω^d, which by Proposition 5.6.2 has the same properties of Ω. □

This function J turns out to be strictly monotone:

Proposition 5.6.4. *Under the same hypotheses of 5.6.3*

$$(J(x) - J(y), x - y) > 0 \tag{5.6.3}$$

for every $x \neq y$

Proof. $(J(x), x) = 1$ by the definition of J and $(J(x), y) < 1$ for every $y \neq x$, $y \in \Omega$ by strict convexity. The same holds of course if the roles of x and y are switched, so that we can conclude

$$(J(x) - J(y), x - y) = 2 - (J(x), y) - (J(y), x) > 0 \tag{5.6.4}$$

□

We now formally state the definition of normal cone for closed convex sets in Banach spaces, which is sometimes stated in terms of subdifferential of the indicator function. As anticipated, it will be a straightforward extension of the definition used for sets in $\mathbb{R}^n$.

Definition 5.6.5. Let Ω be a closed and convex subset of a Banach space X with dual X^*. The normal cone $N_\Omega(\bar{x})$ to Ω in the point $\bar{x} \in \Omega$ is the subset of X^* made by the vectors which supports Ω in $\bar{x}$:

$$N_\Omega(\bar{x}) = \{v \in X^* \mid (v, x - \bar{x}) \leq 0 \ \forall x \in \Omega\} \tag{5.6.5}$$

As for the tangent cones, one can still show generalizing in a straightforward way Proposition 5.1.9 how Clarke's definition applies to convex and closed subsets of Banach spaces (see [2], lemma 3.22 and definition 3.23).
We will instead need a slightly different argument which does not use compactness to prove that $\hat{N}_\Omega(\bar{x}) = T_\Omega(\bar{x})^d$ still holds for Banach spaces.

Proposition 5.6.6. *Let Ω be a closed convex subset of a Banach space X. Then $N_\Omega(\bar{x}) = T_\Omega(\bar{x})^d$ for every point $\bar{x} \in \Omega$.*

Proof. First notice that by the continuity of the dual defining condition

$$C^d = \{v \in X^* \mid (v, x - \bar{x}) \leq 0 \ \forall x \in C\} \tag{5.6.6}$$

we have $C^d = \bar{C}^d$ for any C cone in X.
Let

$$A = \{\lambda(x - \bar{x}) \mid x \in \Omega, \lambda \geq 0\} \tag{5.6.7}$$

so that A is a convex cone in X. Then

$$T_\Omega(\bar{x}) = \mathrm{cl}(\{\lambda(x-\bar{x}) \mid x \in \Omega, \lambda \geq 0\}) = \mathrm{cl}(A) \tag{5.6.8}$$

and

$$\begin{aligned} N_\Omega(\bar{x}) &= \{v \in X^* \mid (v, x-\bar{x}) \leq 0 \; \forall x \in \Omega\} = \\ &= \{v \in X^* \mid (v, \lambda(x-\bar{x})) \leq 0 \; \forall x \in \Omega, \; \lambda \geq 0\} = A^d \end{aligned} \tag{5.6.9}$$

Therefore we can conclude $N_\Omega(\bar{x}) = T_\Omega(\bar{x})^d$ using (5.6.6). □

We can proceed to generalize part of 5.1.10, which is essential to simplify the definition of normalized width.

Proposition 5.6.7. *Let Ω be a closed convex set. For every $r \in X^*$*

$$\max\{0, \sup_{h \in \Omega/\{\bar{x}\}} (r, \frac{h-\bar{x}}{\|h-\bar{x}\|})\} = \mathrm{dist}(r, \hat{N}_\Omega(\bar{x})) \tag{5.6.10}$$

The proof is exactly the same that we presented for $\mathbb{R}^n$.

5.7 NW(Ω) from the dual point of view

The classic FW algorithm for convex optimization on Banach spaces has already been analyzed in [43], where several convergence properties were proved assuming also uniform continuity of the differential.
In the same spirit, here we try to prove convergence properties in a more general setting than $\mathbb{R}^n$ for our FDFW. In the rest of this section X will be a Banach space and Ω will be a convex and bounded subset of X. A few less obvious assumptions will be also needed.
Assumption 1: *X is reflexive.*
Since every closed and convex bounded subset of a Banach space is weakly compact, this assumption guarantees that every linear functional has a maximum and a minimum over Ω. In terms of our algorithms the existence of the FW direction and the in face direction depends on the existence of extreme points for linear continuous functions. In principle we could just assume Ω to be weakly compact. However, together with assumption 2 this would still implies that X is reflexive.
Assumption 2: *Ω has non empty interior.*
This is necessary to guarantee that the width of Ω is greater than 0. It is also not restrictive with respect to assuming that Ω has non empty relative interior. Indeed in this case we can repeat the same analysis translating Ω so that it contains the origin and restricting everything to the vector space aff(Ω). Notice that under these conditions aff(Ω) is a closed subspace of E, hence a reflexive Banach space itself.
Assumption 3a: *The dimension of the faces of Ω is bounded by a finite constant M.*
This is necessary because we need to bound the number of maximal in face steps done by the FDFW as at most a fraction of the total. Indeed at least with the current proof we don't control how much the objective function decreases after one of these

steps.
To avoid a few technical details concerning the definition of faces in Banach spaces we will in practice use the stronger strict convexity assumption.
Assumption 3b: *Ω is strictly convex.*
Under assumptions 1, 2, 3a, we can still define the normalized width for subsets of X generalizing in a straightforward way the normalized width for subsets of $\mathbb{R}^n$. Indeed given $r \in X^*$, $\bar{x} \in \Omega$, we can still define $\mathcal{F}(\bar{x})$ as the minimal face of Ω containing $\bar{x}$, because by assumption all proper faces containing $\bar{x}$ have finite dimension, hence if $x \in \partial\Omega$ their intersection is the proper face with minimal dimension containing $\bar{x}$. The sets $\mathrm{argmin}_{x\in\Omega}(r,x)$, $\mathrm{argmax}_{x\in\mathcal{F}(\bar{x})}(r,x)$ are non empty by assumption 1, so that we can still define $s_r^*(\bar{x})$. Finally, by Proposition 5.6.7 about tangent cones

$$\pi_{\bar{x}}(r) = \max\{0, \sup_{h\in\Omega/\{\bar{x}\}}(r, \frac{h-\bar{x}}{\|h-\bar{x}\|})\} = \mathrm{dist}(r, \hat{N}_\Omega(\bar{x}))$$

so that we have all the elements to define $\mathrm{dirNW}(\Omega, \bar{x}, r)$ and as a consequence $\mathrm{NW}(\Omega)$.
Now under assumptions 1, 2, 3a the linear convergence Theorem 2.5.2 and its corollaries still holds with the same exact proofs. Indeed if f is strongly convex and has Lipschitz differential then it has these properties along every line with the same constants, so that we can write the same inequalities concerning the upper bound on the solution gap and the lower bound on the decrease of the objective function at each step.
We are now interested in giving conditions for $\mathrm{NW}(\Omega)$ to be greater than 0 under assumptions 1, 2, 3b. Without loss of generality we can assume $\{0\} \in \mathrm{int}(\Omega)$, since of course $\mathrm{NW}(\Omega)$ is invariant by translation. First we prove a technical lemma. $\|\cdot\|_*$ will be a norm equivalent to the euclidean norm with dist_* and $B_r^*(x)$ the corresponding distance and ball of center x and radius r.

Lemma 5.7.1. *Let Ω be a convex subset of $\mathbb{R}^2$ such that $0 \in \mathrm{int}(\Omega)$, let and let $r, R > 0$ such that $B_r^*(0) \subset \Omega \subset B_R^*(0)$. Let $c, c^* \in \partial\Omega$. If C^* is the ray generated by c^* then*

$$\mathrm{dist}_*(c, C^*) \geq \frac{1}{1+R/r}\|c-c^*\|_* \tag{5.7.1}$$

Proof. Let $p = (1-t)c^*$ be a projection of p on C^* so that $1-t \geq 0$ and $\mathrm{dist}_*(c, C^*) = \|c-p\|_*$. If $t = 0$ then $p = c^*$ and

$$\|c-c^*\|_* = \mathrm{dist}_*(c, C^*)$$

Otherwise applying a dilatation of center c^* and factor $\frac{1}{t}$ which sends p to the origin we get

$$\|c-p\|_* = |t|\|c^* + \frac{1}{t}(p-c^*) - (c^* + \frac{1}{t}(c-c^*))\|_* = |t|\|c^* + \frac{1}{t}(c-c^*)\|_* \tag{5.7.2}$$

If $0 < t \leq 1$ so that $\frac{1}{t} \geq 1$, since $c^*, c \in \partial\Omega$ we have

$$c^* + \frac{1}{t}(c-c^*) \in (\Omega^\circ)^c \tag{5.7.3}$$

and if $t < 0$ again (5.7.3) holds because $\frac{1}{t} < 0$. Then using the hypothesis $B_r^*(0) \subset \Omega$ we obtain

$$|t|\|c^* + \frac{1}{t}(c - c^*)\|_* \geq |t|r \tag{5.7.4}$$

Concatenating (5.7.2) with (5.7.4), we get

$$\|c - p\|_* \geq |t|r \tag{5.7.5}$$

Now by the definition of p

$$\|c^* - p\|_* = |t|\|c^*\|_* \leq |t|R \tag{5.7.6}$$

and by (5.7.5)

$$\|c^* - p\|_* \leq |t|R = |t|r\frac{R}{r} \leq \|c - p\|_*\frac{R}{r} \tag{5.7.7}$$

From these two inequalities we can finish the proof using the triangular inequality

$$\begin{aligned} \|c - c^*\|_* \leq \|c - p\|_* + \|p - c^*\|_* \leq \|c - p\|_* + |t|R \leq \|c - p\|_*(1 + R/r) = \\ = (1 + R/r)\mathrm{dist}_*(c, C^*) \end{aligned} \tag{5.7.8}$$

□

We use this lemma to compute $\mathrm{dirNW}(\Omega, x, c)$. J will be the generalized duality function introduced in section 5.6; under the hypotheses of Proposition 5.6.3, which are included in the following lemma, J is a bijection between $\partial\Omega$ and $\partial(\Omega^d)$.

Lemma 5.7.2. *Let Ω be a convex and bounded smooth subset of X for which assumptions 1, 2 and 3b hold. Let $r, R > 0$, and assume $0 \in \mathrm{int}(\Omega)$ with $B_r(0) \subset \Omega \subset B_R(0)$. Let $\bar{x} \in \partial\Omega$, $c \in \partial(\Omega^d) \setminus \mathrm{cone}(J(\bar{x}))$, let $x^* = J^{-1}(c)$ and $c^* = J(x)$. Then*

$$\frac{(x^* - \bar{x}, c)}{\|c - c^*\|\|\bar{x} - x^*\|} \leq \mathrm{dirNW}(\Omega, \bar{x}, c) \leq \frac{k(x^* - \bar{x}, c)}{\|c - c^*\|\|x - x^*\|}$$

with $k = 1 + R/r$.

The assumption $0 \in \mathrm{int}(\Omega)$ is not restrictive up to translation.

Proof. Since $c \in \partial\Omega^d$ with Ω smooth and strictly convex $c = J(x^*) \Leftrightarrow c \in N_\Omega(x^*)$. Since X is weakly compact, strictly convex and with non empty interior every linear functional has exactly one solution, so that

$$c \in N_\Omega(x^*) \Rightarrow x^* \in \mathrm{argmax}(c, \cdot) \Rightarrow \{x^*\} = \mathrm{argmax}\{(c, \cdot)\} \tag{5.7.9}$$

Now using the definitions given at the beginning of section 2.4, since by strict convexity $\mathcal{F}(\bar{x}) = \bar{x}$ so that $M_c(\bar{x}) = \bar{x}$, $m_c = \{x^*\}$ by (5.7.9) and

$$\sigma_c^*(\bar{x}) = \inf\{\frac{(c, s - r)}{\|s - \bar{x}\| + \|r - \bar{x}\|} \mid s \in m_c, q \in M_c(\bar{x})\} = \frac{(c, x^* - \bar{x})}{\|\bar{x} - x^*\|} \tag{5.7.10}$$

Therefore

$$\mathrm{dirNW}(\Omega, \bar{x}, c) = \frac{\sigma_c^*(\bar{x})}{\mathrm{dist}_*(c, \mathrm{cone}(c^*))} = \frac{(c, x^* - \bar{x})}{\|\bar{x} - x^*\|\mathrm{dist}_*(c, \mathrm{cone}(c^*))} \tag{5.7.11}$$

the thesis follows if we can prove

$$\|c - c^*\| \leq \text{dist}(c, \text{cone}(c^*)) \leq k\|c - c^*\| \tag{5.7.12}$$

The first inequality is trivial since $c^* \in \text{cone}(c^*)$ and the second follows by applying Lemma 5.7.1 to the plane spanned by c, c^* with the restriction of $\|\cdot\|$ in this plane as norm. $\square$

Theorem 5.7.3. *Under the same hypotheses of Lemma 5.7.2 on* Ω, X *let*

$$M(\Omega) = \inf_{x,y\in\partial\Omega,\ x\neq y} \frac{(x-y, J(x))}{\|x-y\|\|J(x)-J(y)\|} \tag{5.7.13}$$

Then

$$M(\Omega) \geq \text{NW}(\Omega) \geq \min(\frac{r}{D}, \frac{M(\Omega)}{k}) \tag{5.7.14}$$

with $k = 1 + \frac{R}{r}$.

Proof. Since $\text{dirNW}(\Omega, x, c)$ is invariant by positive rescaling of c, we can always assume $c \in \partial(\Omega^d)$ in the rest of the proof. For every $c \in \partial(\Omega^d)$ there exists y_M, y_m such that y_M is the unique maximizer of $(c, \cdot)$ and y_m is the unique minimizer of $(c, \cdot)$. By hypothesis $\pm\hat{c}r \in \Omega$ so that $(c, y_m) \leq (c, -\hat{c}r) = -\|c\|r$ and $(c, y_M) \geq (c, \hat{c}r) = \|c\|r$. Therefore if $x \in \Omega^\circ$ so that $\|\pi_{T_\Omega(x)(c)}\| = \|c\|$

$$\text{dirNW}(\Omega, x, c) = \frac{(y_m - y_M, c)}{\|c\|(\|(y_m - x)\| + \|y_M - x\|)} \geq \frac{2r\|c\|}{2D\|c\|} = \frac{r}{D} \tag{5.7.15}$$

We can now use this result in combination with Lemma 5.7.2 to bound $\text{NW}(\Omega)$:

$$\begin{aligned}
\text{NW}(\Omega) = &\inf_{\substack{x\in\Omega,\\ c\notin N_\Omega(\bar{x})}} \text{dirNW}(\Omega, x, c) = \\
&= \min(\inf_{\substack{x\in\Omega^\circ,\\ c\notin N_\Omega(\bar{x})}} \text{dirNW}(\Omega, x, -c), \inf_{\substack{x\in\partial\Omega,\\ c\notin N_\Omega(\bar{x})}} \text{dirNW}(\Omega, x, c)) \geq \\
&\geq \min(\frac{r}{D}, \inf_{\substack{x\in\partial\Omega,\\ c\notin N_\Omega(\bar{x})}} \frac{(J^{-1}(c) - x, c)}{k\|c - J(x)\|\|x - J^{-1}(c)\|})
\end{aligned} \tag{5.7.16}$$

Now we will use that J is a bijection from $\partial\Omega$ to $\partial(\Omega^d)$ to change variables in the computation of the inf. Setting $y = J^{-1}(c)$ we obtain

$$\inf_{\substack{x\in\partial\Omega,\\ c\notin N_\Omega(\bar{x})}} \frac{(J^{-1}(c) - x, c)}{k\|c - J(x)\|\|x - J^{-1}(c)\|} = \inf_{\substack{x,y\in\partial\Omega,\\ x\neq y}} \frac{(y - x, c)}{k\|J(y) - J(x)\|\|x - y\|} \tag{5.7.17}$$

Applying this last equation to (5.7.16) we get the desired lower bound on $\text{NW}(\Omega)$. The upper bound can be proved exactly in the same way, ignoring the case $x \in \Omega^\circ$. $\square$

Corollary 5.7.4. *Under the same hypotheses Theorem* (5.7.3)

$$\inf_{\substack{x,y\in\partial\Omega,\\ x\neq y}} \frac{(x - y, J(x) - J(y))}{2\|x - y\|\|J(x) - J(y)\|} \geq \text{NW}(\Omega) \tag{5.7.18}$$

Proof. Follows immediately from the upper bound in Theorem 5.7.3 above switching the roles of x and y, summing and using the superadditivity of the inf operator. $\square$

Remark 5.7.5. Notice that in this upper bound for NW(Ω) the numerator is exactly the measure of monotonicity of J, which is sort of normalized dividing by the norms of the two terms in the denominator. In $\mathbb{R}^n$, if θ is the angle between $x - y$ and $J(x) - J(y)$, then the value of the upper bound is $\cos(\theta)/2$.

Remark 5.7.6. We now explain how the upper bound

$$u(x, y) = \frac{(x - y, J(x) - J(y))}{2\|x - y\|\|J(x) - J(y)\|} \tag{5.7.19}$$

is related to the positive definiteness of $DJ(x)$. Assume that $\partial\Omega \times \partial\Omega$ is strongly compact. The upper bound $u(x, y)$ is continuous and strictly positive in the set $\{(x, y) \in \partial\Omega \mid x \neq y\}$. Then it is bounded away from 0 if and only if $u(x_k, y_k) \to c > 0$ for every $(x_k, y_k) \to (\bar{x}, \bar{x})$, $\bar{x} \in \Omega$. If J is regular enough this is equivalent to say

$$\lim_{y_k \to \bar{x}} \frac{(\bar{x} - y_k, DJ(\bar{x})(\bar{x} - y_k))}{\|\bar{x} - y_k\|\|DJ(\bar{x})(\bar{x} - y_k)\|} > 0 \tag{5.7.20}$$

for every y_k sequence in the tangent space to Ω Finally, inequality (5.7.20) holds if $DJ(\bar{x})$ is positive definite and has continuous inverse.

Bibliography

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

[2] Viorel Barbu and Teodor Precupanu. *Convexity and optimization in Banach spaces*. Springer Science & Business Media, 2012.

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

[4] Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[5] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

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

[7] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

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

[9] Immanuel M. Bomze, Francesco Rinaldi, and Samuel Rota Bulo'. First-order methods for the impatient: support identification in finite time with convergent Frank-Wolfe variants. *eprints for the optimization community*, 2018.

[10] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

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

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

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

[14] Qiong Cao, Yiming Ying, and Peng Li. Distance metric learning revisited. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 283–298. Springer, 2012.

[15] Coralia Cartis and Nicholas IM Gould. Finding a point in the relative interior of a polyhedron. 2007.

[16] Katherine Christianson. Notes on fulton, section 1.2 convex polyhedral cones. 2016.

[17] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. Integer programming, volume 271 of graduate texts in mathematics, 2014.

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

[19] JC Dunn and E Sachs. The effect of perturbations on the convergence rates of optimization algorithms. *Applied Mathematics and Optimization*, 10(1):143–157, 1983.

[20] Joseph C Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

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

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

[23] Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *ICML*, volume 15, pages 541–549, 2015.

[24] FJ Gould and Jon W Tolle. A necessary and sufficient qualification for constrained optimization. *SIAM Journal on Applied Mathematics*, 20(2):164–172, 1971.

[25] Luigi Grippo and Marco Sciandrone. *Metodi di ottimizzazione non vincolata.* Springer Science & Business Media, 2011.

[26] Peter M Gruber. Approximation of convex bodies. In *Convexity and its Applications*, pages 131–162. Springer, 1983.

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

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

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

[30] Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting frank-wolfe. *arXiv preprint arXiv:1810.02429*, 2018.

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

[32] Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[33] Simon Lacoste-Julien and Martin Jaggi. An affine invariant linear convergence analysis for frank-wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

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

[35] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate frank-wolfe optimization for structural svms. *arXiv preprint arXiv:1207.4747*, 2012.

[36] Guoyin Li. Global error bounds for piecewise convex polynomials. *Mathematical Programming*, 137(1-2):37–64, 2013.

[37] Sanjay Mehrotra and Yinyu Ye. Finding an interior point in the optimal face of linear programs. *Mathematical Programming*, 62(1-3):497–515, 1993.

[38] Julie Nutini, Mark Schmidt, and Warren Hare. " active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *arXiv preprint arXiv:1712.03577*, 2017.

[39] Javier Pena and Daniel Rodriguez. Polytope conditioning and linear convergence of the frank–wolfe algorithm. *Mathematics of Operations Research*, 2018.

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

[41] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

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

[43] Hong-Kun Xu. Convergence analysis of the frank-wolfe algorithm and its generalization in banach spaces. *arXiv preprint arXiv:1710.07367*, 2017.

[44] Yi Xu and Tianbao Yang. Frank-wolfe method is automatically adaptive to error bound condition. *arXiv preprint arXiv:1810.04765*, 2018.