

---

# An Affine Invariant Linear Convergence Analysis for Frank-Wolfe Algorithms

---

**Simon Lacoste-Julien**  
INRIA - SIERRA project-team  
École Normale Supérieure, Paris, France

**Martin Jaggi**  
Simons Institute  
UC Berkeley, USA

## Abstract

We study the linear convergence of variants of the Frank-Wolfe algorithms for some classes of strongly convex problems, using only affine-invariant quantities. As in [GM86], we show the linear convergence of the standard Frank-Wolfe algorithm when the solution is in the interior of the domain, but with affine invariant constants. We also show the linear convergence of the away-steps variant of the Frank-Wolfe algorithm, but with constants which only depend on the geometry of the domain, and not any property of the location of the optimal solution. Running these algorithms does not require knowing any problem specific parameters.

The Frank-Wolfe algorithm [FW56] (also known as *conditional gradient*) is one of the earliest existing methods for constrained convex optimization, and has seen an impressive revival recently due to its nice properties compared to projected or proximal gradient methods, in particular for sparse optimization and machine learning applications.

On the other hand, the classical projected gradient and proximal methods have been known to exhibit a very nice adaptive acceleration property, namely that the convergence rate becomes linear for strongly convex objective, i.e. that the optimization error of the same algorithm after  $k$  iterations will decrease geometrically with  $O(\rho^{-k})$  instead of the usual  $O(1/k)$  for general convex objective functions. It has become an active research topic recently whether such an acceleration is also possible for Frank-Wolfe type methods.

**Contributions.** We show that the Frank-Wolfe algorithm with away-steps converges linearly (i.e. with a geometric rate) for any strongly convex objective function optimized over a polytope domain, with a constant bounded away from zero that only depends on the geometry of the polytope. Our convergence analysis is affine invariant (both the algorithm and the convergence rate are unaffected by an affine transformation of the variables). Also, our analysis does not depend on the location of the true optimum with respect to the domain, which was a disadvantage of earlier existing results such as [Wol70, GM86, BT04], and the later results [AST08, KY10, AFNS13] that need Robinson’s condition [Rob82]. Our analysis yields a weaker sufficient condition than Robinson’s condition; in particular we can have linear convergence even in some cases when the function has more than one global minima and is not globally strongly convex. As a second contribution, we provide an affine invariant version of the analysis of [GM86] showing that the classical (unmodified) Frank-Wolfe algorithm converges linearly on strongly convex functions when the optimum lies in the interior.

**Related Work.** The away-steps variant of the Frank-Wolfe algorithm, that can also remove weight from “bad” ones of the currently active atoms, was proposed in [Wol70], and later also analyzed in [GM86]. The precise algorithm is stated below in Algorithm 1. An alternative away-step algorithm (with a sublinear convergence rate) has been considered by [Cla10], namely performing an away step whenever the number of atoms of non-zero weight has exceeded a fixed target size. The disadvantage of this method is that it requires knowledge of the curvature constant, which is not realistic in many practical applications. For the classical Frank-Wolfe algorithm, the early work of [LP66, Theorem 6.1] has shown a linear convergence rate under the strong requirement that the objective is strongly convex, and furthermore the domain is strongly convex as a set. [BT04] has shown a

linear rate for the special case of quadratic objectives when the optimum is in the strict interior of the domain, but their result was already subsumed by [GM86]. More recently [AST08, KY10, AFÑS13] have obtained linear convergence results in the case that the optimum solution satisfies Robinson’s condition [Rob82]. In a different recent line of work, [GH13a, GH13b] has studied an algorithm variation<sup>1</sup> that moves mass from the worst vertices to the “towards” vertex until a specific condition is satisfied, yielding a linear convergence rate. Their algorithm requires the knowledge of several constants though, and moreover is not adaptive to the best-case scenario, unlike the Frank-Wolfe algorithm with away steps and line-search. None of these previous works was shown to be affine invariant, and most require additional knowledge about problem specific parameters.

## 1 Frank-Wolfe Algorithms, and Away-Steps

We consider general constrained convex optimization problems of the form

$$\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}).$$

We assume  $f$  is convex and differentiable, and that the domain  $\mathcal{D}$  is a bounded convex subset of a vector space. The Frank-Wolfe method [FW56], also known as *conditional gradient* [LP66] works as follows: At a current  $\mathbf{x}^{(k)}$ , the algorithm considers the linearization of the objective function, and moves slightly towards a minimizer of this linear function (taken over the same domain). In terms of convergence, it is known that the iterates of Frank-Wolfe satisfy  $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq O(1/k)$ , for  $\mathbf{x}^*$  being an optimal solution [FW56, DH78, Jag13]. One of the main reasons for the recent increased popularity of Frank-Wolfe-type algorithms is the sparsity of the iterates, i.e. that the iterate is always represented as a sparse convex combination of at most  $k$  vertices  $\mathcal{S}^{(k)} \subseteq \mathcal{V}$  of the domain  $\mathcal{D}$ , which we write as  $\mathbf{x}^{(k)} = \sum_{\mathbf{v} \in \mathcal{S}^{(k)}} \alpha_{\mathbf{v}}^{(k)} \mathbf{v}$ . Here  $\mathcal{V}$  is defined to be the set of vertices (extreme points) of  $\mathcal{D}$ , so that  $\mathcal{D} = \text{conv}(\mathcal{V})$ . We assume that the *linear oracle* defining  $\mathbf{s}_k$  always returns a point from  $\mathcal{V}$  as a minimizer.

**Away-Steps.** The away-steps variant of Frank-Wolfe, as stated in Algorithm 1, was proposed in [Wol70], with the idea to also remove weight from “bad” ones of the currently active atoms. Note that the classical Frank-Wolfe algorithm is obtained by only using the FW direction in Algorithm 1. If  $\gamma_k = \gamma_{\max}$ , then we call this step a *drop step*, as it fully removes the vertex  $\mathbf{v}_k$  from the currently active set of atoms  $\mathcal{S}^{(k)}$ . The updates of the algorithm are of the following form: For a FW step, we have  $\mathcal{S}^{(k+1)} = \{\mathbf{s}_k\}$  if  $\gamma_k = 1$ ; otherwise  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \cup \{\mathbf{s}_k\}$ . Also, we have  $\alpha_{\mathbf{s}_k}^{(k+1)} := (1 - \gamma_k)\alpha_{\mathbf{s}_k}^{(k)} + \gamma_k$  and  $\alpha_{\mathbf{v}}^{(k+1)} := (1 - \gamma_k)\alpha_{\mathbf{v}}^{(k)}$  for  $\mathbf{v} \in \mathcal{S}^{(k)} \setminus \{\mathbf{s}_k\}$ . For an away step, we have  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \setminus \{\mathbf{v}_k\}$  if  $\gamma_k = \gamma_{\max}$  (a *drop step*); otherwise  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)}$ . Also, we have  $\alpha_{\mathbf{v}_k}^{(k+1)} := (1 + \gamma_k)\alpha_{\mathbf{v}_k}^{(k)} - \gamma_k$  and  $\alpha_{\mathbf{v}}^{(k+1)} := (1 + \gamma_k)\alpha_{\mathbf{v}}^{(k)}$  for  $\mathbf{v} \in \mathcal{S}^{(k)} \setminus \{\mathbf{v}_k\}$ .

**Away-Steps.** The away-steps variant of Frank-Wolfe, as stated in Algorithm 1, was proposed in [Wol70], with the idea to also remove weight from “bad” ones of the currently active atoms. Note that the classical Frank-Wolfe algorithm is obtained by only using the FW direction in Algorithm 1. If  $\gamma_k = \gamma_{\max}$ , then we call this step a *drop step*, as it fully removes the vertex  $\mathbf{v}_k$  from the currently active set of atoms  $\mathcal{S}^{(k)}$ . The updates of the algorithm are of the following form: For a FW step, we have  $\mathcal{S}^{(k+1)} = \{\mathbf{s}_k\}$  if  $\gamma_k = 1$ ; otherwise  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \cup \{\mathbf{s}_k\}$ . Also, we have  $\alpha_{\mathbf{s}_k}^{(k+1)} := (1 - \gamma_k)\alpha_{\mathbf{s}_k}^{(k)} + \gamma_k$  and  $\alpha_{\mathbf{v}}^{(k+1)} := (1 - \gamma_k)\alpha_{\mathbf{v}}^{(k)}$  for  $\mathbf{v} \in \mathcal{S}^{(k)} \setminus \{\mathbf{s}_k\}$ . For an away step, we have  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \setminus \{\mathbf{v}_k\}$  if  $\gamma_k = \gamma_{\max}$  (a *drop step*); otherwise  $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)}$ . Also, we have  $\alpha_{\mathbf{v}_k}^{(k+1)} := (1 + \gamma_k)\alpha_{\mathbf{v}_k}^{(k)} - \gamma_k$  and  $\alpha_{\mathbf{v}}^{(k+1)} := (1 + \gamma_k)\alpha_{\mathbf{v}}^{(k)}$  for  $\mathbf{v} \in \mathcal{S}^{(k)} \setminus \{\mathbf{v}_k\}$ .

## 2 Affine Invariant Measures of Smoothness and Strong Convexity

**Affine Invariance.** An optimization method is called *affine invariant* if it is invariant under affine transformations of the input problem: If one chooses any re-parameterization of the domain  $\mathcal{D}$ , by a *surjective* linear or affine map  $M : \hat{\mathcal{D}} \rightarrow \mathcal{D}$ , then the “old” and “new” optimization problems  $\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x})$  and  $\min_{\hat{\mathbf{x}} \in \hat{\mathcal{D}}} \hat{f}(\hat{\mathbf{x}})$  for  $\hat{f}(\hat{\mathbf{x}}) := f(M\hat{\mathbf{x}})$  look completely the same to the algorithm.

<sup>1</sup>This can be interpreted as a concrete instantiation of the stronger oracle proposed in [Lan13].

|  |
|--|
| <p><b>Algorithm 1:</b> Frank-Wolfe Algorithm with Away Steps</p> <p>Let <math>\mathbf{x}^{(0)} \in \mathcal{V}</math>, and <math>\mathcal{S}^{(0)} := \{\mathbf{x}^{(0)}\}</math><br/> <small>(so that <math>\alpha_{\mathbf{v}}^{(0)} = 1</math> for <math>\mathbf{v} = \mathbf{x}^{(0)}</math> and 0 otherwise)</small></p> <p><b>for</b> <math>k = 0 \dots K</math> <b>do</b></p> <p style="padding-left: 2em;">Let <math>\mathbf{s}_k \in \arg \min_{\mathbf{v} \in \mathcal{V}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{v} \rangle</math> and <math>\mathbf{d}_k^{\text{FW}} := \mathbf{s}_k - \mathbf{x}^{(k)}</math><br/> <small>(the FW direction)</small></p> <p style="padding-left: 2em;">Let <math>\mathbf{v}_k \in \arg \max_{\mathbf{v} \in \mathcal{S}^{(k)}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{v} \rangle</math> and <math>\mathbf{d}_k^{\text{A}} := \mathbf{x}^{(k)} - \mathbf{v}_k</math><br/> <small>(the away direction)</small></p> <p style="padding-left: 2em;"><b>if</b> <math>\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k^{\text{FW}} \rangle \leq \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k^{\text{A}} \rangle</math> <b>then</b></p> <p style="padding-left: 4em;"><math>\mathbf{d}_k := \mathbf{d}_k^{\text{FW}}</math>, and <math>\gamma_{\max} := 1</math> <small>(choose the FW direction)</small></p> <p style="padding-left: 2em;"><b>else</b></p> <p style="padding-left: 4em;"><math>\mathbf{d}_k := \mathbf{d}_k^{\text{A}}</math>, and <math>\gamma_{\max} := \alpha_{\mathbf{v}_k} / (1 - \alpha_{\mathbf{v}_k})</math><br/> <small>(choose the away direction, and maximum feasible step-size)</small></p> <p style="padding-left: 2em;"><b>end</b></p> <p style="padding-left: 2em;">Line search: <math>\gamma_k \in \arg \min_{\gamma \in [0, \gamma_{\max}]} f(\mathbf{x}^{(k)} + \gamma \mathbf{d}_k)</math></p> <p style="padding-left: 2em;">Update <math>\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \gamma_k \mathbf{d}_k</math><br/> <small>(and accordingly for the weights <math>\alpha^{(k+1)}</math>, see text)</small></p> <p style="padding-left: 2em;">Update <math>\mathcal{S}^{(k+1)} := \{\mathbf{v} \text{ s.t. } \alpha_{\mathbf{v}}^{(k+1)} &gt; 0\}</math></p> <p><b>end</b></p> |
|--|

More precisely, every “new” iterate must remain exactly the transform of the corresponding old iterate; an affine invariant analysis should thus yield the convergence rate and constants unchanged by the transformation. It is well known that Newton’s method is affine invariant under invertible  $M$ , and the Frank-Wolfe algorithm is affine invariant in the even stronger sense under arbitrary  $M$  [Jag13]. (This is directly implied if the algorithm and all constants appearing in the analysis only depend on inner products with the gradient, which are preserved since  $\nabla \hat{f} = M^T \nabla f$ .)

**Affine Invariant Measures of Smoothness.** The affine invariant convergence analysis of the standard Frank-Wolfe algorithm by [Jag13] crucially relies on the following measure of non-linearity of the objective function  $f$  over the domain  $\mathcal{D}$ . The *curvature constant*  $C_f$  of a convex and differentiable function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , with respect to a compact domain  $\mathcal{D}$  is defined as

$$C_f := \sup_{\substack{\mathbf{x}, \mathbf{s} \in \mathcal{D}, \gamma \in [0,1], \\ \mathbf{y} = \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle). \quad (1)$$

The assumption of bounded curvature  $C_f$  closely corresponds to a Lipschitz assumption on the gradient of  $f$ . More precisely, if  $\nabla f$  is  $L$ -Lipschitz continuous on  $\mathcal{D}$  with respect to some arbitrary chosen norm  $\|\cdot\|$ , then

$$C_f \leq \text{diam}_{\|\cdot\|}(\mathcal{D})^2 L, \quad (2)$$

where  $\text{diam}_{\|\cdot\|}(\cdot)$  denotes the  $\|\cdot\|$ -diameter, see [Jag13, Lemma 7]. While the early papers [FW56, Dun79] on the Frank-Wolfe algorithm relied on such Lipschitz constants with respect to a norm, the curvature constant  $C_f$  here is affine invariant, does not depend on any norm, and gives tighter convergence rates.  $C_f$  combines the complexity of  $\mathcal{D}$  and the curvature of  $f$  into a single quantity.

**An Affine Invariant Notion of Strong Convexity.** Inspired by the affine invariant curvature measure, one can also define a related affine invariant measure of strong convexity, when combined with the assumption of the optimum  $\mathbf{x}^*$  being in the strict interior of  $\mathcal{D}$ :

$$\mu_f^{\text{FW}} := \inf_{\substack{\mathbf{x} \in \mathcal{D} \setminus \{\mathbf{x}^*\}, \gamma \in (0,1), \\ \bar{\mathbf{s}} = \bar{\mathbf{s}}(\mathbf{x}, \mathbf{x}^*, \mathcal{D}), \\ \mathbf{y} = \mathbf{x} + \gamma(\bar{\mathbf{s}} - \mathbf{x})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle). \quad (3)$$

Here the point  $\bar{\mathbf{s}}$  is defined to be the point where the ray from  $\mathbf{x}$  to the optimum  $\mathbf{x}^*$  pinches the boundary of the set  $\mathcal{D}$ , i.e. furthest away from  $\mathbf{x}$  while still in  $\mathcal{D}$ ,  $\bar{\mathbf{s}}(\mathbf{x}, \mathbf{x}^*, \mathcal{D}) := \text{ray}(\mathbf{x}, \mathbf{x}^*) \cap \partial\mathcal{D}$ . We will later show that this strict interior assumption, which can be very prohibitive, can be removed for the Frank-Wolfe algorithm with *away steps*, as we explain in Section 4. Clearly, the quantity  $\mu_f^{\text{FW}}$  is affine invariant, as it only depends on the inner products of feasible points with the gradient.

**Remark 1.** For all pairs of functions  $f$  and bounded sets  $\mathcal{D}$ , it holds that  $\mu_f^{\text{FW}} \leq C_f$ .

The following simple lemma gives an interpretation of the very abstract (affine invariant) quantity defined above, in terms of classical norms and strong-convexity properties.

**Lemma 2.** Let  $f$  be a convex differentiable function and suppose  $f$  is strongly convex w.r.t. some arbitrary norm  $\|\cdot\|$  over the domain  $\mathcal{D}$  with strong-convexity constant  $\mu > 0$ . Furthermore, suppose that the (unique) optimum  $\mathbf{x}^*$  lies in the relative interior of  $\mathcal{D}$ , i.e.  $\delta_{\mathbf{x}^*, \mathcal{D}} := \inf_{\mathbf{s} \in \partial\mathcal{D}} \|\mathbf{s} - \mathbf{x}^*\| > 0$ . Then

$$\mu_f^{\text{FW}} \geq \mu \cdot \delta_{\mathbf{x}^*, \mathcal{D}}^2.$$

### 3 Linear Convergence of Frank-Wolfe

We obtain an affine invariant linear convergence proof for the standard FW algorithm when  $f$  is strongly convex and the solution  $\mathbf{x}^*$  lies in the relative interior of  $\mathcal{D}$  (an improvement over [GM86]).

**Theorem 3.** Suppose that  $f$  has smoothness constant  $C_f$  as defined in (1), as well as “interior” strong convexity constant  $\mu_f^{\text{FW}}$  as defined in (3). Then the error of the iterates of the Frank-Wolfe algorithm with step-size  $\gamma := \min\{1, \frac{g_k}{C_f}\}$  (or using line-search) decreases geometrically, that is

$$h_{k+1} \leq \left(1 - \rho_f^{\text{FW}}\right) h_k,$$

where  $\rho_f^{\text{FW}} := \min\{\frac{1}{2}, \frac{\mu_f^{\text{FW}}}{C_f}\}$ . Here in each iteration,  $h_k := f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)$  denotes the primal error, and  $g_k := g(\mathbf{x}^{(k)}) := \max_{\mathbf{s} \in \mathcal{D}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{s} \rangle$  is the duality gap as defined by [Jag13].

### 4 Linear Convergence of Frank-Wolfe with Away-Steps

We now show the linear convergence of FW with away-steps under strong convexity, without any assumption on the location of the optimum with respect to the domain. However, our convergence rate will depend on a purely geometric complexity constant of the domain  $\mathcal{D}$ , as we show below.

**An Affine Invariant Notion of Strong Convexity which Depends on the Geometry of  $\mathcal{D}$ .** The trick is to use anchor points in the domain in order to define standard lengths (by looking at proportions on lines). These anchor points ( $\mathbf{s}_f(\mathbf{x})$  and  $\mathbf{v}_f(\mathbf{x})$  defined below) are motivated directly from the away-steps algorithm.

Let  $\mathbf{s}_f(\mathbf{x}) := \arg \min_{\mathbf{v} \in \mathcal{V}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$  (the standard Frank-Wolfe direction). To define the away-vertex, we consider all possible expansions of  $\mathbf{x}$  as a convex combination of vertices. Let  $\mathcal{S}_x := \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{V} \text{ such that } \mathbf{x} \text{ is a proper}^2 \text{ convex combination of all the elements in } \mathcal{S}\}$ . For a given set  $\mathcal{S}$ , we write  $\mathbf{v}_S(\mathbf{x}) := \arg \max_{\mathbf{v} \in \mathcal{S}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$  for the away vertex in the algorithm supposing that the current set of active vertices was  $\mathcal{S}$ . Finally, we define  $\mathbf{v}_f(\mathbf{x}) := \arg \min_{\{\mathbf{v} = \mathbf{v}_S(\mathbf{x}) \mid \mathcal{S} \in \mathcal{S}_x\}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$  to be

the worst-case away vertex (that is, the vertex which would yield the smallest away descent).

We can now define the strong convexity constant  $\mu_f^A$  which depends *both* on the function  $f$  and the domain  $\mathcal{D}$ :

$$\mu_f^A := \inf_{\mathbf{x} \in \mathcal{D}} \inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle < 0}} \frac{2}{\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle). \quad (4)$$

Here the positive quantity  $\gamma^A(\mathbf{x}, \mathbf{x}^*) := \frac{\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle}{\langle \nabla f(\mathbf{x}), \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}$  plays the role of  $\gamma$  in definition (1).

**Interpretation.** The above complexity definition is already sufficient for us to prove the linear convergence. Additionally, the constant can be understood in terms of the geometry of  $\mathcal{D}$ , as follows:

**Directional Width.** The directional width of a set  $\mathcal{D}$  with respect to a direction  $\mathbf{d}$  (and underlying inner product norm  $\|\cdot\|$ ) is defined as  $\text{dir}W(\mathcal{D}, \mathbf{d}) := \max_{\mathbf{x} \in \mathcal{D}} \langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{x} \rangle - \min_{\mathbf{x} \in \mathcal{D}} \langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{x} \rangle$ .

**Pyramidal Width.** We define the pyramidal directional width of a set  $\mathcal{D}$  with respect to a direction  $\mathbf{d}$  and a base point  $\mathbf{x} \in \mathcal{D}$  to be  $\text{Pdir}W(\mathcal{D}, \mathbf{d}, \mathbf{x}) := \min_{\mathcal{S} \in \mathcal{S}_x} \text{dir}W(\mathcal{S} \cup \{\mathbf{s}(\mathcal{D}, \mathbf{d})\}, \mathbf{d})$  where  $\mathbf{s}(\mathcal{D}, \mathbf{d}) := \arg \max_{\mathbf{v} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{v} \rangle$ . To define the pyramidal width of a set, we take the infimum over a set of possible feasible directions  $\mathbf{d}$  (in order to avoid the problem of zero width). A direction  $\mathbf{d}$  is *feasible* for  $\mathcal{D}$  from  $\mathbf{x}$  if it points inwards the set, (i.e.  $\mathbf{d} \in \text{cone}(\mathcal{D} - \mathbf{x})$ ).

We define the *pyramidal width* of a set  $\mathcal{D}$  to be the smallest pyramidal width of all its faces, i.e.

$$\text{Pdir}W(\mathcal{D}) := \inf_{\substack{\mathcal{K} \in \text{faces}(\mathcal{D}) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{d} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{\mathbf{0}\}}} \text{Pdir}W(\mathcal{K}, \mathbf{d}, \mathbf{x}). \quad (5)$$

**Remark 4.** Any curved domain will yield a pyramidal width of zero, because then the set of active atoms  $\mathcal{S}_x$  can contain vertices arbitrary close to the boundary forming a very narrow pyramid. The pyramidal width quantity is thus only useful on polytopes (the convex hull of a finite set of points).

**Remark 5.** Let  $\mathbf{v}(\mathbf{x}, \mathbf{d}) :=$  the vertex which achieves the minimum in  $\min_{\mathcal{S} \in \mathcal{S}_x} \max_{\mathbf{v} \in \mathcal{S}} \langle \mathbf{d}, -\mathbf{v} \rangle$  for a polytope  $\mathcal{K}$  and  $\mathbf{x} \in \mathcal{K}$ . Then we have  $\text{Pdir}W(\mathcal{K}, \mathbf{d}, \mathbf{x}) = \langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{s}(\mathcal{K}, \mathbf{d}) - \mathbf{v}(\mathbf{x}, \mathbf{d}) \rangle$ .

We conjecture that  $\text{Pdir}W(\mathcal{D})$  for  $\mathcal{D}$  being the unit simplex in  $\mathbb{R}^d$  is  $\frac{2}{\sqrt{d}}$ .

**Lemma 6.** Let  $f$  be a convex differentiable function and suppose that  $f$  is  $\mu$ -strongly convex w.r.t. some inner product norm  $\|\cdot\|$  over the domain  $\mathcal{D}$  with strong-convexity constant  $\mu \geq 0$ . Then

$$\mu_f^A \geq \mu \cdot (\text{Pdir}W(\mathcal{D}))^2.$$

**Theorem 7.** Suppose that  $f$  has smoothness constant  $C_f^A$ ,<sup>3</sup> as well as geometric strong convexity constant  $\mu_f^A$  as defined in (4). Then the error of the iterates of the FW algorithm with away-steps<sup>4</sup> (Algorithm 1) decreases geometrically at each step that is not a drop step (i.e. when  $\gamma_k < \gamma_{\max}$ ), that is

$$h_{k+1} \leq \left(1 - \rho_f^A\right) h_k,$$

where  $\rho_f^A := \frac{\mu_f^A}{4C_f^A}$ . Moreover, the number of drop steps up to iteration  $k$  is bounded by  $k/2$ . This yields the global linear convergence rate of  $h_k \leq h_0 \exp(-\frac{1}{2}\rho_f^A k)$ .

<sup>2</sup>By *proper* convex combination, we mean that all coefficients are non-zero in the convex combination.

<sup>3</sup>For a convenience in the proof, we use a slightly modified curvature constant  $C_f^A$ , which is identical to the definition of  $C_f$ , except that both positive and negative step-sizes are allowed, i.e. the range of  $\gamma$  in the definition for  $C_f$  is replaced by  $[-1, 1]$  instead of just  $[0, 1]$ . Note that boundedness of this (again affine invariant)  $C_f^A$  is still implied by the Lipschitz continuity of the gradient of  $f$  (over the slightly larger domain  $\mathcal{D} + \mathcal{D} - \mathcal{D}$ , but with the same diameter constant).

<sup>4</sup>In the algorithm, one can either use line-search or set the step-size as the feasible one that minimizes the quadratic upper bound given by the curvature  $C_f^A$ , i.e.  $\gamma_k := \min\{1, \gamma_{\max}, \gamma_k^B\}$  where  $\gamma_k^B := \frac{g_k}{2C_f^A}$  and  $g_k := \langle -\nabla f(\mathbf{x}^{(k)}), \mathbf{s}_k - \mathbf{v}_k \rangle$ .

**Acknowledgements.** Simon Lacoste-Julien acknowledges support by the ERC (SIERRA-ERC-239993). Martin Jaggi acknowledges support by the Simons Institute for the Theory of Computing, by the Swiss National Science Foundation (SNSF), and by ERC Project SIPA.

## References

- [AFÑS13] Hector Allende, Emanuele Frandi, Ricardo Nanculef, and Claudio Sartori. [Novel Frank-Wolfe Methods for SVM Learning](#). *arXiv.org*, 2013.
- [AST08] Selin Damla Ahipaaoğlu, Peng Sun, and Michael Todd. [Linear Convergence of a Modified Frank-Wolfe Algorithm for Computing Minimum-Volume Enclosing Ellipsoids](#). *Optimization Methods and Software*, 23(1):5–19, 2008.
- [BT04] Amir Beck and Marc Teboulle. [A Conditional Gradient Method with Linear Rate of Convergence for Solving Convex Linear Systems](#). *Mathematical Methods of Operations Research (ZOR)*, 59(2):235–247, 2004.
- [Cla10] Kenneth L Clarkson. [Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm](#). *ACM Transactions on Algorithms*, 6(4), 2010.
- [DH78] Joseph C Dunn and S Harshbarger. [Conditional Gradient Algorithms with Open Loop Step Size Rules](#). *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.
- [Dun79] 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.
- [FG13] Robert M Freund and Paul Grigas. [New Analysis and Results for the Conditional Gradient Method](#). *arXiv.org*, 2013.
- [FW56] Marguerite Frank and Philip Wolfe. [An Algorithm for Quadratic Programming](#). *Naval Research Logistics Quarterly*, 3:95–110, 1956.
- [GH13a] Dan Garber and Elad Hazan. [A Linearly Convergent Conditional Gradient Algorithm with Applications to Online and Stochastic Optimization](#). *arXiv.org*, 2013.
- [GH13b] Dan Garber and Elad Hazan. [Playing Non-linear Games with Linear Oracles](#). *FOCS 2013 - 54th Annual Symposium on Foundations of Computer Science*, 420–428, 2013.
- [GM86] Jacques Guélat and Patrice Marcotte. [Some Comments on Wolfe’s ‘Away Step’](#). *Mathematical Programming*, 35(1):110–119, 1986.
- [Jag11] Martin Jaggi. [Sparse Convex Optimization Methods for Machine Learning](#). PhD thesis, ETH Zürich, 2011.
- [Jag13] Martin Jaggi. [Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization](#). In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [KY10] Piyush Kumar and E Alper Yildirim. [A Linearly Convergent Linear-Time First-Order Algorithm for Support Vector Classification with a Core Set Result](#). *INFORMS Journal on Computing*, 2010.
- [Lan13] Guanghui Lan. [The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle](#). *optimization-online.org*, 2013.
- [LJJS13] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. [Block-Coordinate Frank-Wolfe Optimization for Structural SVMs](#). In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [LP66] Evgenij S Levitin and Boris T Polyak. [Constrained Minimization Methods](#). *USSR Computational Mathematics and Mathematical Physics*, 6(5):787–823, 1966.
- [Rob82] Stephen M Robinson. [Generalized Equations and their Solutions, Part II: Applications to Nonlinear Programming](#). Springer Berlin Heidelberg, 1982.
- [Wol70] Philip Wolfe. [Convergence Theory in Nonlinear Programming](#). In J Abadie, editor, *Integer and Nonlinear Programming*, pages 1–23. North-Holland, 1970.
- [Zie95] Günter M Ziegler. [Lectures on Polytopes](#), volume 152 of *Graduate Texts in Mathematics*. Springer Verlag, 1995.

## A Linear Convergence of Frank-Wolfe for Strongly Convex Functions with Optimum in the Interior

### A.1 An Affine Invariant Notion of Strong Convexity

We re-state and interpret the ‘‘interior’’ strong convexity constant  $\mu_f^{\text{FW}}$  as defined in (3), that is

$$\mu_f^{\text{FW}} := \inf_{\substack{\mathbf{x} \in \mathcal{D} \setminus \{\mathbf{x}^*\}, \\ \bar{\mathbf{s}} = \bar{\mathbf{s}}(\mathbf{x}, \mathbf{x}^*, \mathcal{D}), \\ \gamma \in (0, 1], \\ \mathbf{y} = \mathbf{x} + \gamma(\bar{\mathbf{s}} - \mathbf{x})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle).$$

Here the point  $\bar{\mathbf{s}}$  is defined to be the point where the ray from  $\mathbf{x}$  to  $\mathbf{x}^*$  pinches the boundary of the set  $\mathcal{D}$ , i.e.  $\bar{\mathbf{s}}(\mathbf{x}, \mathbf{x}^*, \mathcal{D}) := \text{ray}(\mathbf{x}, \mathbf{x}^*) \cap \partial\mathcal{D}$ .

Recalling that the curvature  $C_f$  by definition (1) provides an affine-invariant quadratic upper bound on the function  $f$ , the strong convexity constant  $\mu_f^{\text{FW}}$  here gives rise to an analogous quadratic lower bound, that is

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{\gamma^2}{2} \mu_f^{\text{FW}} \quad (6)$$

if the point  $\mathbf{y} = \mathbf{x} + \gamma(\bar{\mathbf{s}} - \mathbf{x})$  is determined by the boundary point  $\bar{\mathbf{s}}(\mathbf{x}, \mathbf{x}^*, \mathcal{D})$  and an arbitrary step-size  $\gamma \in [0, 1]$ , i.e., if the point  $\mathbf{y}$  lies on the segment which is between  $\mathbf{x}$  and the boundary of  $\mathcal{D}$ , and passes through  $\mathbf{x}^*$ .

Here we prove Lemma 2, which gives a simple geometric interpretation of the abstract (affine-invariant) quantity  $\mu_f^{\text{FW}}$  defined above, in terms of classical norms and strong-convexity properties.

**Lemma’ 2.** *Let  $f$  be a convex differentiable function and suppose  $f$  is strongly convex w.r.t. some arbitrary norm  $\|\cdot\|$  over the domain  $\mathcal{D}$  with strong-convexity constant  $\mu > 0$ .*

*Furthermore, suppose that the (unique) optimum  $\mathbf{x}^*$  lies in the relative interior of  $\mathcal{D}$ , i.e.  $\delta_{\mathbf{x}^*, \mathcal{D}} := \inf_{\mathbf{s} \in \partial\mathcal{D}} \|\mathbf{s} - \mathbf{x}^*\| > 0$ . Then*

$$\mu_f^{\text{FW}} \geq \mu \cdot \delta_{\mathbf{x}^*, \mathcal{D}}^2.$$

*Proof.* By definition of strong convexity with respect to a norm, we have that for any  $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ ,

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

We want to use this lower bound in the definition (3) of the affine invariant strong convexity constant. Observe that  $\frac{1}{\gamma^2} \|\mathbf{y} - \mathbf{x}\|^2 = \|\bar{\mathbf{s}} - \mathbf{x}\|^2$  for any  $\mathbf{x}$  used in (3) since  $\mathbf{y} := \mathbf{x} + \gamma(\bar{\mathbf{s}} - \mathbf{x}) \in \mathcal{D}$  by convexity. Moreover, by the definition of  $\bar{\mathbf{s}}$  and  $\delta_{\mathbf{x}^*, \mathcal{D}}$ ,  $\|\bar{\mathbf{s}} - \mathbf{x}\| \geq \|\bar{\mathbf{s}} - \mathbf{x}^*\| \geq \delta_{\mathbf{x}^*, \mathcal{D}}$ . Therefore, we can lower bound  $\mu_f^{\text{FW}}$  as

$$\mu_f^{\text{FW}} \geq \inf \frac{2}{\gamma^2} \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 = \inf \mu \|\bar{\mathbf{s}} - \mathbf{x}\|^2 \geq \mu \cdot \delta_{\mathbf{x}^*, \mathcal{D}}^2,$$

which is the claimed bound.  $\square$

### A.2 Convergence Analysis

**Curvature.** The definition of the curvature constant  $C_f$  as in (1) directly gives an affine invariant quadratic upper bound on the objective function, as follows:

Let  $\mathbf{x}_\gamma := \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})$  be the point obtained by moving with step-size  $\gamma$  in direction  $\mathbf{s} \in \mathcal{D}$ . By definition of  $C_f$ , we have

$$f(\mathbf{x}_\gamma) \leq f(\mathbf{x}) + \gamma \langle \nabla f(\mathbf{x}), \mathbf{s} - \mathbf{x} \rangle + \frac{\gamma^2}{2} C_f, \quad \forall \gamma \in [0, 1].$$

This crucial bound enables us to analyze the objective improvement in each iteration in Frank-Wolfe-type algorithms, as in [Jag13]: If the point  $\mathbf{s}$  is the standard Frank-Wolfe direction returned by an exact linear oracle, then the middle quantity is exactly the negative of the duality gap,  $\langle \nabla f(\mathbf{x}), \mathbf{s} - \mathbf{x} \rangle = -g(\mathbf{x})$ . If an inexact linear oracle is used instead, which has multiplicative approximation quality  $\nu$  (to be defined below), then we always have the upper bound

$$f(\mathbf{x}_\gamma) \leq f(\mathbf{x}) - \gamma \nu g(\mathbf{x}) + \frac{\gamma^2}{2} C_f, \quad \forall \gamma \in [0, 1]. \quad (7)$$

**Inexact Linear Oracles.** The standard linear oracle used inside the classical Frank-Wolfe algorithm is given by  $\mathbf{s} \in \arg \min_{\mathbf{v} \in \mathcal{D}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ . We say that the linear oracle satisfies multiplicative accuracy  $\nu$  for some  $\nu \in [0, 1]$ , if for any  $\mathbf{x} \in \mathcal{D}$ , the returned  $\mathbf{s}$  is such that

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{s} \rangle \geq \nu \cdot \max_{\mathbf{s}' \in \mathcal{D}} \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{s}' \rangle . \quad (8)$$

Note that the classical Frank-Wolfe direction  $\mathbf{s}$  satisfies this inequality with  $\nu = 1$ . The inequality means that the oracle answer  $\mathbf{s}$  attains at least a  $\nu$ -fraction of the current *duality gap*  $g(\mathbf{x}) := \max_{\mathbf{s} \in \mathcal{D}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{s} \rangle$  as defined by [Jag13].

*Related work.* The sublinear convergence of Frank-Wolfe with  $O(1/k)$  is known to also hold if this linear subproblems are only solved approximately (meaning that the linear oracle is inexact). For additive approximation accuracy, this was shown by [DH78, Dun79] for the line-search case, and by [Jag11, Jag13] for the simpler  $\frac{2}{k+2}$  step-size and the primal-dual convergence. For multiplicative accuracy (relative to the duality gap), it was shown by [LJSP13, Appendix C]. The case of the inexact or noisy gradient information can also be analyzed in the same way, as discussed in [Jag13, FG13].

**Linear Convergence Proof.** Here we prove a slightly stronger version of Theorem 3, showing the linear convergence also in the case where the linear subproblems in each iteration are only solved approximately. The exact oracle case is obtained for  $\nu := 1$ .

**Theorem' 3.** *Suppose that  $f$  has smoothness constant  $C_f$  as defined in (1), as well as “interior” strong convexity constant  $\mu_f^{\text{FW}}$  as defined in (3).*

*Then the error of the iterates of the Frank-Wolfe algorithm with step-size  $\gamma := \min\{1, \frac{\nu g_k}{C_f}\}$  (or using line-search) decreases geometrically, that is*

$$h_{k+1} \leq \left(1 - \rho_f^{\text{FW}}\right) h_k ,$$

where  $\rho_f^{\text{FW}} := \min\{\frac{\nu}{2}, \nu^2 \frac{\mu_f^{\text{FW}}}{C_f}\}$ . Here in each iteration,  $h_k := f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)$  denotes the primal error, and  $g_k := g(\mathbf{x}^{(k)}) := \max_{\mathbf{s} \in \mathcal{D}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{s} \rangle$  is the duality gap as defined by [Jag13], and  $\nu \in [0, 1]$  is the multiplicative approximation quality to which the linear sub-problems are solved.

*Proof.* Applying the strong convexity bound (6) at the current iterate  $\mathbf{x} := \mathbf{x}^{(k)}$  for the special step-size  $\bar{\gamma}$  such that  $\mathbf{y} = \mathbf{x}^{(k)} + \bar{\gamma}(\bar{\mathbf{s}} - \mathbf{x}^{(k)}) = \mathbf{x}^*$  gives

$$\begin{aligned} \frac{\bar{\gamma}^2}{2} \mu_f^{\text{FW}} &\leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ &= f(\mathbf{x}^*) - f(\mathbf{x}^{(k)}) - \bar{\gamma} \langle \nabla f(\mathbf{x}^{(k)}), \bar{\mathbf{s}} - \mathbf{x}^{(k)} \rangle \\ &\leq -h_k + \bar{\gamma} g_k . \end{aligned}$$

Therefore  $h_k \leq -\frac{\bar{\gamma}^2}{2} \mu_f^{\text{FW}} + \bar{\gamma} g_k$ , which is upper bounded by  $\frac{g_k^2}{2\mu_f^{\text{FW}}}$ .

(Here we have used the trivial inequality  $0 \leq a^2 - 2ab + b^2$  for the choice of numbers  $a := \frac{g_k}{\mu_f^{\text{FW}}}$  and  $b := \bar{\gamma}$ )

We now want to use the curvature definition to lower bound the absolute progress  $h_k - h_{k+1}$ . The definition of the curvature  $C_f$  in the form of the quadratic upper bound (7) reads as  $h_k - h_{k+1} \geq \gamma \nu g_k - \frac{\gamma^2}{2} C_f$ . Using this for the particular step-size  $\gamma := \frac{\nu g_k}{C_f}$ , the r.h.s. is  $\frac{\nu^2 g_k^2}{2C_f}$ . (The border case when  $\frac{\nu g_k}{C_f} > 1$  will be discussed separately below). The same inequality also holds in the line-search case, as the improvement only gets better. Combining the two bounds, we have obtained

$$\frac{h_k - h_{k+1}}{h_k} \geq \nu^2 \frac{\mu_f^{\text{FW}}}{C_f}$$

implying that we have a geometric rate of decrease  $h_{k+1} \leq \left(1 - \nu^2 \frac{\mu_f^{\text{FW}}}{C_f}\right) h_k$ .

*Border case.* In the above analysis, we have assumed that the step-size  $\gamma := \frac{\nu g_k}{C_f} \leq 1$ . If this is not the case (i.e. if  $\nu g_k > C_f$ ), then the actual step-size in the algorithm is clipped to 1, in which case the curvature upper bound (7) for  $\gamma := 1$  gives  $h_k - h_{k+1} \geq \nu g_k - \frac{1}{2} C_f > \nu g_k - \frac{1}{2} \nu g_k = \frac{\nu}{2} g_k$ . Using that the main property  $h_k \leq g_k$  of the duality gap (by convexity), we therefore have  $\frac{h_k - h_{k+1}}{h_k} > \frac{\nu}{2}$ , which gives a geometric decrease of the error with constant  $1 - \frac{\nu}{2}$ .  $\square$

## B Linear Convergence of FW with Away-Steps under Strong Convexity

### B.1 Interpretation of the Geometric Strong Convexity Constant $\mu_f^A$

The geometric strong convexity constant  $\mu_f^A$ , as defined in (4), is affine invariant, since it only depends on the inner products of feasible points with the gradient. Also, it combines both the complexity of the function  $f$  and the geometry of the domain  $\mathcal{D}$ . The goal of this subsection is to prove Lemma 6, which provides a geometric interpretation of  $\mu_f^A$ . The lemma allows us to bound the constant  $\mu_f^A$  in terms of the strong convexity of the objective function, combined with a purely geometric complexity measure of the domain  $\mathcal{D}$ . In the following Section B.2 below, we will show the linear convergence of Algorithm 1 under the assumption that  $\mu_f^A > 0$ . From the view of Lemma 6,  $\mu_f^A > 0$  is a slightly weaker condition than the strong convexity of the function over a polytope domain (it is implied by strong convexity).

We recall the definition of  $\mu_f^A$  as given in (4):

$$\mu_f^A := \inf_{\mathbf{x} \in \mathcal{D}} \inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle < 0}} \frac{2}{\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle).$$

Here the positive quantity  $\gamma^A(\mathbf{x}, \mathbf{x}^*) := \frac{\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle}{\langle \nabla f(\mathbf{x}), \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}$  plays the role of  $\gamma$  in the analogous upper bound definition (1) for the curvature. We recall that  $\mathbf{s}_f(\mathbf{x}) := \arg \min_{\mathbf{v} \in \mathcal{V}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$  and that  $\mathbf{v}_f(\mathbf{x}) := \arg \min_{\{\mathbf{v} = \mathbf{v}_S(\mathbf{x}) \mid S \in \mathcal{S}_{\mathbf{x}}\}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ .

We recall the definition of the pyramidal directional width of a set  $\mathcal{D}$  with respect to a direction  $\mathbf{d}$  and a base point  $\mathbf{x} \in \mathcal{D}$ :  $PdirW(\mathcal{D}, \mathbf{d}, \mathbf{x}) := \min_{S \in \mathcal{S}_{\mathbf{x}}} dirW(S \cup \{\mathbf{s}(\mathcal{D}, \mathbf{d})\}, \mathbf{d})$  where  $\mathbf{s}(\mathcal{D}, \mathbf{d}) := \arg \max_{\mathbf{v} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{v} \rangle$ . We now provide a proof of Remark 5, which will be useful at the end of the proof of Lemma 6.

**Remark' 5.** Let  $\mathbf{v}(\mathbf{x}, \mathbf{d}) :=$  the vertex which achieves the minimizer of  $\min_{S \in \mathcal{S}_{\mathbf{x}}} \max_{\mathbf{v} \in S} \langle \mathbf{d}, -\mathbf{v} \rangle$  for a polytope  $\mathcal{K}$  and  $\mathbf{x} \in \mathcal{K}$ . Then we have

$$PdirW(\mathcal{K}, \mathbf{d}, \mathbf{x}) = \left\langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{s}(\mathcal{K}, \mathbf{d}) - \mathbf{v}(\mathbf{x}, \mathbf{d}) \right\rangle. \quad (9)$$

*Proof.*

$$\begin{aligned} PdirW(\mathcal{K}, \mathbf{d}, \mathbf{x}) &= \frac{1}{\|\mathbf{d}\|_*} \min_{S \in \mathcal{S}_{\mathbf{x}}} \left( \max_{\mathbf{y} \in S \cup \{\mathbf{s}(\mathcal{K}, \mathbf{d})\}} \langle \mathbf{d}, \mathbf{y} \rangle - \min_{\mathbf{y} \in S \cup \{\mathbf{s}(\mathcal{K}, \mathbf{d})\}} \langle \mathbf{d}, \mathbf{y} \rangle \right) \\ &= \frac{1}{\|\mathbf{d}\|_*} \min_{S \in \mathcal{S}_{\mathbf{x}}} \left( \langle \mathbf{d}, \mathbf{s}(\mathcal{K}, \mathbf{d}) \rangle + \max_{\mathbf{y} \in S} \langle \mathbf{d}, -\mathbf{y} \rangle \right) \\ &= \frac{1}{\|\mathbf{d}\|_*} \left( \langle \mathbf{d}, \mathbf{s}(\mathcal{K}, \mathbf{d}) \rangle + \min_{S \in \mathcal{S}_{\mathbf{x}}} \max_{\mathbf{y} \in S} \langle \mathbf{d}, -\mathbf{y} \rangle \right) \\ &= \left\langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{s}(\mathcal{K}, \mathbf{d}) - \mathbf{v}(\mathbf{x}, \mathbf{d}) \right\rangle. \end{aligned}$$

□

**Exposing a facet of a polytope.** Finally, we introduce a final concept that will be useful in the proof. We say that a direction  $\mathbf{d}$  **exposes a facet**<sup>5</sup>  $\mathcal{F}$  of the polytope  $\mathcal{D}$  at  $\mathbf{x}$  if 1)  $\mathcal{F}$  includes  $\mathbf{x}$  and is a facet of  $\mathcal{D}$ ; and 2) the orthogonal component of  $\mathbf{d}$  to this facet defines this facet with  $\mathbf{d}$  on one side and  $\mathcal{D} - \mathbf{x}$  on the other side. In other words, let  $\mathcal{F}_s := \text{span}(\mathcal{D} - \mathbf{x})$  be the affine hull of  $\mathcal{F}$

<sup>5</sup>As a reminder, we define a  $k$ -face of  $\mathcal{D}$  (a  $k$ -dimensional face of  $\mathcal{D}$ ) a set  $\mathcal{K}$  such that  $\mathcal{K} = \mathcal{D} \cap \{\mathbf{y} : \langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle = 0\}$  for some normal vector  $\mathbf{r}$  and fixed reference point  $\mathbf{x} \in \mathcal{K}$  with the additional property that  $\mathcal{D}$  lies on one side of the given half-space determined by  $\mathbf{r}$  i.e.  $\langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle \leq 0 \forall \mathbf{y} \in \mathcal{D}$ .  $k$  is the dimensionality of the affine hull of  $\mathcal{K}$ . We call a  $k$ -face of dimensions  $k = 0, 1, \dim(\mathcal{D}) - 2$  and  $\dim(\mathcal{D}) - 1$  a *vertex*, *edge*, *ridge* and *facet* respectively.  $\mathcal{D}$  is a  $k$ -face of itself with  $k = \dim(\mathcal{D})$ . See definition 2.1 in [Zie95].

re-centered at  $\mathbf{x}$ ; let  $\mathbf{P}_{\mathcal{F}_s}$  be the orthogonal projection operator onto  $\mathcal{F}_s$ ; then the second condition can be expressed as  $\mathcal{F} = \{\mathbf{y} \in \mathcal{D} : \langle (\mathbf{I} - \mathbf{P}_{\mathcal{F}_s})\mathbf{d}, \mathbf{y} - \mathbf{x} \rangle = 0\}$  and  $\langle (\mathbf{I} - \mathbf{P}_{\mathcal{F}_s})\mathbf{d}, \mathbf{y} - \mathbf{x} \rangle \leq 0 \forall \mathbf{y} \in \mathcal{D}$  (note that  $(\mathbf{I} - \mathbf{P}_{\mathcal{F}_s})\mathbf{d}$  is the orthogonal component of  $\mathbf{d}$  to the facet  $\mathcal{F}$ ). Note that these conditions imply that  $\mathbf{d}$  cannot be a feasible direction, i.e.  $\mathbf{d} \notin \text{cone}(\mathcal{D} - \mathbf{x})$  and that  $\mathbf{x}$  must be on the (relative) boundary of  $\mathcal{D}$ . It turns out that the converse is also true: if  $\mathbf{d} \notin \text{cone}(\mathcal{D} - \mathbf{x})$ , then there must exist at least a facet of  $\mathcal{D}$  exposed by  $\mathbf{d}$  at  $\mathbf{x}$ .<sup>6</sup>

**Lemma' 6.** *Let  $f$  be a convex differentiable function and suppose that  $f$  is  $\mu$ -strongly convex w.r.t. some inner product norm  $\|\cdot\|$  over the domain  $\mathcal{D}$  with strong-convexity constant  $\mu \geq 0$ . Then*

$$\mu_f^A \geq \mu \cdot (\text{Pdir}W(\mathcal{D}))^2 .$$

*Proof.* By definition of strong convexity with respect to a norm, we have that for any  $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ ,

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 . \quad (10)$$

Using the strong convexity bound (10) with  $\mathbf{y} := \mathbf{x}^*$  on the right hand side of equation (4) (and using the shorthand  $\mathbf{r}_x := -\nabla f(\mathbf{x})$ ), we thus get:

$$\begin{aligned} \mu_f^A &\geq \inf_{\substack{\mathbf{x}, \mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_x, \mathbf{x}^* - \mathbf{x} \rangle > 0}} \mu \left( \frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \mathbf{x}^* - \mathbf{x} \rangle} \|\mathbf{x}^* - \mathbf{x}\| \right)^2 \\ &= \mu \inf_{\substack{\mathbf{x} \neq \mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle > 0}} \left( \frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} \right)^2 , \end{aligned} \quad (11)$$

where  $\hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} := \frac{\mathbf{x}^* - \mathbf{x}}{\|\mathbf{x}^* - \mathbf{x}\|}$  is the unit norm feasible direction from  $\mathbf{x}$  to  $\mathbf{x}^*$ . We are thus taking an infimum over all possible feasible directions starting from  $\mathbf{x}$  (i.e. which moves within  $\mathcal{D}$ ) with the additional constraint that it makes a positive inner product with the negative gradient  $\mathbf{r}_x$  i.e. it is a strict descent direction. This is only possible if  $\mathbf{x}$  is not already optimal, i.e.  $\mathbf{x} \in \mathcal{D} \setminus \mathcal{X}^*$  where  $\mathcal{X}^* := \{\mathbf{x}^* \in \mathcal{D} : \langle \mathbf{r}_{\mathbf{x}^*}, \mathbf{x} - \mathbf{x}^* \rangle \leq 0 \forall \mathbf{x} \in \mathcal{D}\}$  is the set of optimal points. [NOTE: I know that by strong convexity it only contains one point; but I wanted to keep it general here just to see the effect of the constraints and to get more intuition about the constants].

The goal in the rest of the proof is to equivalently project  $\mathbf{r}_x$  onto facets of  $\mathcal{D}$  and then to characterize the property of its projection so that we can consider a wider set of valid directions that will thus yield a lower bound on the infimum of  $\frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle}$ . For the rest of the proof, we fix  $\mathbf{x} \notin \mathcal{X}^*$

and we work on the centered polytope at  $\mathbf{x}$  i.e. let  $\tilde{\mathcal{D}} = \mathcal{D} - \mathbf{x}$ . During the proof, we work on faces  $\mathcal{K}_l$  of  $\tilde{\mathcal{D}}$  of decreasing dimensions which all include  $\mathbf{x}$  at their origin, as well as maintain a projection of the gradient  $\mathbf{d}_l \in \mathcal{C}_l := \text{span}(\mathcal{K}_l)$ . We let  $\mathbf{P}_l$  be the orthogonal projection operator onto  $\mathcal{C}_l$ . We will keep projecting the gradient as  $\mathbf{d}_l := \mathbf{P}_l \mathbf{d}_{l-1}$  until they become a non-zero feasible direction from the origin i.e.  $\mathbf{d}_l \in \text{cone}(\mathcal{K}_l) \setminus \{\mathbf{0}\}$ , at which point we will exit the loop with  $\mathbf{d} = \mathbf{d}_l$  and  $\mathcal{K} = \mathcal{K}_l$  for the last considered face.

We start with  $\mathcal{K}_0 = \tilde{\mathcal{D}}$  and we note that since both  $\mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x})$  and  $\hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*}$  belong to  $\mathcal{C}_0 = \text{span}(\mathcal{K}_0)$ , if we let  $\mathbf{d}_0 = \mathbf{P}_0 \mathbf{r}_x$ , then we have  $\frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} = \frac{\langle \mathbf{d}_0, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{d}_0, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle}$  for any  $\mathbf{x}^*$  such that  $\langle \mathbf{r}_x, \mathbf{x}^* - \mathbf{x} \rangle \neq 0$ . Then we consider whether  $\mathbf{d}_0$  is a feasible direction in  $\mathcal{K}_0$ . If  $\mathbf{d}_0$  is feasible i.e.  $\mathbf{d}_0 \in \text{cone}(\mathcal{K}_0)$ , then we stop with  $\mathbf{d} = \mathbf{d}_0 = \mathbf{P}_0 \mathbf{r}_x$  and  $\mathcal{K} = \mathcal{K}_0$ . By the definition of the dual norm  $\|\cdot\|_*$  (generalized Cauchy-Schwartz), we have  $\langle \mathbf{d}, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle \leq \|\mathbf{d}\|_* \|\hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*}\| = \|\mathbf{d}\|_* \cdot 1$ , and thus for this  $\mathbf{x}$  we have:

$$\inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle > 0}} \frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} \geq \left\langle \frac{\mathbf{d}_0}{\|\mathbf{d}_0\|_*}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \right\rangle .$$

In the other possibility ( $\mathbf{d}_0 \notin \text{cone}(\mathcal{K}_0)$ ), then there must exist a least one facet  $\mathcal{K}_1$  of  $\mathcal{K}_0$  that is exposed by  $\mathbf{d}_0$  at  $\mathbf{0}$  (note that we cannot have  $\mathbf{d}_0 = \mathbf{0}$  since  $\mathbf{x} \notin \mathcal{X}^*$ ). We now project  $\mathbf{d}_0$  on

<sup>6</sup>To find such an exposed facet, consider the  $\mathcal{H}$ -polyhedron representation of  $\text{cone}(\mathcal{D} - \mathbf{x})$  (see [Zie95]). As  $\mathbf{d}$  is not feasible, at least one halfspace constraint must be violated; the intersection of the hyperplane determining this halfspace constraint with  $\mathcal{D} - \mathbf{x}$  yields (the translation of) one exposed facet.

$\text{span}(\mathcal{K}_1)$ :  $\mathbf{d}_1 := \mathbf{P}_1 \mathbf{d}_0$ , and we show how the lower bound transforms. This yields the following inequalities:

$$\begin{aligned}
\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle &= \max_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{r}_x, \mathbf{s} - \mathbf{x} \rangle + \min_{S \in \mathcal{S}_x} \max_{\mathbf{v} \in S} \langle -\mathbf{r}_x, \mathbf{v} - \mathbf{x} \rangle \\
&= \max_{\mathbf{y} \in \mathcal{K}_0} \langle \mathbf{d}_0, \mathbf{y} \rangle + \min_{S \in \mathcal{S}_x} \max_{\mathbf{v} \in S} \langle -\mathbf{d}_0, \mathbf{v} - \mathbf{x} \rangle \\
&\geq \max_{\mathbf{y} \in \mathcal{K}_1} \langle \mathbf{d}_0, \mathbf{y} \rangle + \min_{S \in \mathcal{S}_x} \max_{\mathbf{v} \in S \cap (\mathcal{K}_1 + \mathbf{x})} \langle -\mathbf{d}_0, \mathbf{v} - \mathbf{x} \rangle \\
&= \max_{\mathbf{y} \in \mathcal{K}_1} \langle \mathbf{d}_1, \mathbf{y} \rangle + \min_{S \in \mathcal{S}_x} \max_{\mathbf{v} \in S} \langle -\mathbf{d}_1, \mathbf{v} - \mathbf{x} \rangle \\
&= \langle \mathbf{d}_1, \mathbf{s}(\mathcal{K}_1, \mathbf{d}_1) \rangle + \langle -\mathbf{d}_1, \mathbf{v}(\mathbf{x}, \mathbf{d}_1) - \mathbf{x} \rangle. \tag{12}
\end{aligned}$$

From the first to the second line, we used the fact that  $\langle \mathbf{r}_x - \mathbf{d}_0, \mathbf{y} \rangle = 0$  for any  $\mathbf{y} \in \mathcal{K}_0 = \mathcal{D} - \mathbf{x}$  as  $\mathbf{d}_0$  is the orthogonal projection of  $\mathbf{r}_x$  on  $\mathcal{C}_0 = \text{span}(\mathcal{K}_0)$  (and thus we also have that  $\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{x} \rangle = \langle \mathbf{d}_0, \mathbf{s}(\mathcal{K}_0, \mathbf{d}_0) \rangle$ ). To go from the second to the third line, we use the fact that the first term yields an inequality as  $\mathcal{K}_1 \subseteq \mathcal{K}_0$ . Also, let  $\mathcal{K}_x$  be the minimal dimensional face of  $\mathcal{D}$  containing  $\mathbf{x}$  (and thus  $\mathbf{x}$  is in the relative interior of  $\mathcal{K}_x$ ). Note that  $\bigcup \mathcal{S}_x = \text{vertices}(\mathcal{K}_x)$ , and also that  $\mathcal{K}_x$  is included in any other face containing  $\mathbf{x}$ . We thus have  $S \subseteq \mathcal{K}_1 + \mathbf{x}$  for any  $S \in \mathcal{S}_x$  and thus the second term on the second line yielded an equality. The fourth line used the fact that  $\mathbf{d}_0 - \mathbf{d}_1$  is orthogonal to members of  $\mathcal{K}_1$ . The fifth line used the definition of  $\mathbf{s}(\mathcal{K}_1, \mathbf{d}_1)$  and introduced the notation  $\mathbf{v}(\mathbf{x}, \mathbf{d}) :=$  the vertex  $\mathbf{v} \in \mathcal{K}_x$  which achieves the minimizer of  $\min_{S \in \mathcal{S}_x} \max_{\mathbf{v} \in S} \langle \mathbf{d}, -\mathbf{v} \rangle$ .

To deal with  $\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle = \langle \mathbf{d}_0, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle$ , we use the crucial fact that  $\mathbf{d}_0$  exposes the facet  $\mathcal{K}_1$  of  $\mathcal{K}_0$ . This implies that  $\langle \mathbf{d}_0 - \mathbf{P}_0 \mathbf{d}_0, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle \leq 0$  for all  $\mathbf{x}^* - \mathbf{x} \in \mathcal{K}_0 \setminus \{\mathbf{0}\}$ . So consider  $\mathbf{r}_0 := \arg \max_{\substack{\mathbf{y} \in \mathcal{K}_0 \\ \langle \mathbf{d}_0, \mathbf{y} \rangle > 0}} \left\langle \mathbf{d}_0, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle$ . We claim that we can choose  $\mathbf{r}_0 \in \mathcal{K}_1$ . To see this, let  $\mathbf{r}_1 = \mathbf{P}_1 \mathbf{r}_0$  and write  $\mathbf{r}_1^\perp = \mathbf{r}_0 - \mathbf{r}_1$  and  $\mathbf{d}_1^\perp = \mathbf{d}_0 - \mathbf{d}_1$ . Then we have:

$$\begin{aligned}
\left\langle \mathbf{d}_0, \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|} \right\rangle &= \frac{1}{\|\mathbf{r}_0\|} \langle \mathbf{d}_1 + \mathbf{d}_1^\perp, \mathbf{r}_1 + \mathbf{r}_1^\perp \rangle \\
&= \frac{1}{\|\mathbf{r}_0\|} (\langle \mathbf{d}_1, \mathbf{r}_1 \rangle + 0 + \underbrace{\langle \mathbf{d}_1^\perp, \mathbf{r}_1 + \mathbf{r}_1^\perp \rangle}_{\leq 0}) \\
&\leq \frac{1}{\|\mathbf{r}_0\|} \langle \mathbf{d}_1, \mathbf{r}_1 \rangle \leq \frac{1}{\|\mathbf{r}_1\|} \langle \mathbf{d}_1, \mathbf{r}_1 \rangle, \text{ and thus,} \\
\max_{\substack{\mathbf{y} \in \mathcal{K}_0 \\ \langle \mathbf{d}_0, \mathbf{y} \rangle > 0}} \left\langle \mathbf{d}_0, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle &= \max_{\substack{\mathbf{y} \in \mathcal{K}_1 \\ \langle \mathbf{d}_1, \mathbf{y} \rangle > 0}} \left\langle \mathbf{d}_1, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle. \tag{13}
\end{aligned}$$

Note that in the third line, we have used that  $\|\mathbf{r}_1\| = \|\mathbf{P}_1 \mathbf{r}_0 - \mathbf{P}_1 \mathbf{0}\| \leq \|\mathbf{r}_0 - \mathbf{0}\|$  by the contraction property of the orthogonal projection for inner product norms.<sup>7</sup> In the last line, we have an equality instead of the  $\leq$  inequality as  $\langle \mathbf{d}_1, \mathbf{y} \rangle = \langle \mathbf{d}_0, \mathbf{y} \rangle \forall \mathbf{y} \in \mathcal{K}_1$  and  $\mathcal{K}_1 \subseteq \mathcal{K}_0$ , and so we also have the  $\geq$  direction. Combining the facts from (12) and (13), we get in this case:

$$\inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle > 0}} \frac{\langle \mathbf{r}_x, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_x, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} \geq \langle \mathbf{d}_1, \mathbf{s}(\mathcal{K}_1, \mathbf{d}_1) + \mathbf{x} - \mathbf{v}(\mathbf{x}, \mathbf{d}_1) \rangle \left( \max_{\substack{\mathbf{y} \in \mathcal{K}_1 \\ \langle \mathbf{d}_1, \mathbf{y} \rangle > 0}} \left\langle \mathbf{d}_1, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle \right)^{-1}$$

We are now back to a similar situation as before, but with  $\mathcal{K}_1$  instead of  $\mathcal{K}_0$  as the reference polytope. Note that by the third line of (13), we have  $\langle \mathbf{d}_1, \mathbf{r}_1 \rangle \geq \langle \mathbf{d}_0, \mathbf{r}_0 \rangle > 0$  and thus  $\mathbf{d}_1 \neq \mathbf{0}$  (which is crucial to avoid a trivial lower bound of zero). So again, we consider whether  $\mathbf{d}_1 \in \text{cone}(\mathcal{K}_1)$ . If  $\mathbf{d}_1 \in \text{cone}(\mathcal{K}_1)$ , we stop here with  $\mathbf{d} = \mathbf{d}_1$  and  $\mathcal{K} = \mathcal{K}_1$ . By Cauchy-Schwartz, we again have

<sup>7</sup>The contraction property is only valid for inner product norms (i.e.  $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ ), so this is where the assumption that the norm was generated by an inner product comes into play.

$\max_{\substack{\mathbf{y} \in \mathcal{K} \\ \langle \mathbf{d}, \mathbf{y} \rangle > 0}} \left\langle \mathbf{d}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle \leq \|\mathbf{d}\|_*$ , and so we conclude

$$\inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_{\mathbf{x}}, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle > 0}} \frac{\langle \mathbf{r}_{\mathbf{x}}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_{\mathbf{x}}, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} \geq \left\langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{s}(\mathcal{K} + \mathbf{x}, \mathbf{d}) - \mathbf{v}(\mathbf{x}, \mathbf{d}) \right\rangle \quad (14)$$

where  $\mathbf{d} \in \text{cone}(\mathcal{K}) \setminus \{\mathbf{0}\}$ .

If  $\mathbf{d}_1 \notin \text{cone}(\mathcal{K}_1)$ , then we continue our iterative process: we get that  $\mathbf{d}_1$  exposes a facet  $\mathcal{K}_2$  of  $\mathcal{K}_1$ . We thus project  $\mathbf{d}_1$  on  $\mathcal{K}_2$  to get  $\mathbf{d}_2 = \mathbf{P}_2 \mathbf{d}_1$ . We can repeat exactly the same argument as before to get (12) and (13) with  $\mathbf{d}_2$  and  $\mathcal{K}_2$  in place of  $\mathbf{d}_1$  and  $\mathcal{K}_1$  (and  $\mathbf{d}_2 \neq \mathbf{0}$ ). If  $\mathbf{d}_2 \in \text{cone}(\mathcal{K}_2)$ , then we stop with  $\mathbf{d} = \mathbf{d}_2$  and  $\mathcal{K} = \mathcal{K}_2$  and we again get the inequality (14). Otherwise, we get an exposed facet  $\mathcal{K}_3$ , and repeat the process with  $\mathbf{d}_3 = \mathbf{P}_3 \mathbf{d}_2$ . This process must stop at some point  $l$ : at the latest, we will reach  $\mathcal{K}_l = \mathcal{K}_{\mathbf{x}} - \mathbf{x}$ , the minimal dimensional face containing  $\mathbf{0}$ . In this case we must have  $\mathbf{d}_l \in \text{cone}(\mathcal{K}_l)$  as  $\mathbf{0}$  is in the relative interior of  $\mathcal{K}_l$  for a minimal face and so all directions are feasible. We also note that  $\mathbf{d}_l \neq \mathbf{0}$  by the argument in (13) that implies  $\langle \mathbf{d}_l, \mathbf{s}(\mathcal{K}_l, \mathbf{d}_l) \rangle > 0$  (this condition is crucial to avoid having a lower bound of zero!). The latter also implies that the dimensionality of  $\mathcal{K}_l$  must at least be 1. Letting again  $\mathbf{d} = \mathbf{d}_l$  and  $\mathcal{K} = \mathcal{K}_l$ , we get inequality (14) with  $\mathbf{d} \in \text{cone}(\mathcal{K}) \setminus \{\mathbf{0}\}$ .

From this argument, we can see that by considering all the possible faces of  $\tilde{\mathcal{D}}$  of dimension at least one which includes  $\mathbf{0}$ , and any feasible directions for these faces, we are sure to include the  $\mathbf{d}$  and  $\mathcal{K}$  that appears in (14). Translating back to the affine space  $\mathcal{D}$  (i.e. we use  $\mathcal{K} + \mathbf{x}$  as the face of  $\mathcal{D}$  which contains  $\mathbf{x}$ ), we can start to vary  $\mathbf{x}$  again. We thus obtain the following lower bound:

$$\begin{aligned} \inf_{\mathbf{x} \notin \mathcal{X}^*} \inf_{\substack{\mathbf{x}^* \in \mathcal{D} \\ \text{s.t. } \langle \mathbf{r}_{\mathbf{x}}, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle > 0}} \frac{\langle \mathbf{r}_{\mathbf{x}}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_{\mathbf{x}}, \hat{\mathbf{r}}_{\mathbf{x}, \mathbf{x}^*} \rangle} &\geq \inf_{\mathbf{x} \notin \mathcal{X}^*} \inf_{\substack{\mathcal{K} \in \text{faces}(\mathcal{D}) \\ \mathcal{K} \ni \mathbf{x} \\ \mathbf{d} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{\mathbf{0}\}}} \left\langle \frac{\mathbf{d}}{\|\mathbf{d}\|_*}, \mathbf{s}(\mathcal{K}, \mathbf{d}) - \mathbf{v}(\mathbf{x}, \mathbf{d}) \right\rangle \\ &\geq \inf_{\substack{\mathcal{K} \in \text{faces}(\mathcal{D}) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{d} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{\mathbf{0}\}}} \text{Pdir}W(\mathcal{K}, \mathbf{d}, \mathbf{x}) = \text{Pdir}W(\mathcal{D}). \end{aligned}$$

For the last inequality, we used (9) from Remark 5. Combining this statement with (11) concludes the proof.  $\square$

## B.2 Linear Convergence Proof

**Curvature Constants.** Because of the additional possibility of the away step in Algorithm 1, we need to define the following slightly modified additional curvature constant, which will be needed for the linear convergence analysis of the algorithm<sup>8</sup>:

$$C_f^- := \sup_{\substack{\mathbf{x}, \mathbf{s} \in \mathcal{D}, \\ \gamma \in [0, 1], \\ \mathbf{y} = \mathbf{x} + \gamma(\mathbf{x} - \mathbf{s})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle). \quad (15)$$

By comparing with  $C_f$  (1), we see that the modification is that  $\mathbf{y}$  is defined with the *away* direction  $\mathbf{x} - \mathbf{s}$  instead of a standard FW direction  $\mathbf{s} - \mathbf{x}$ . This might yield some  $\mathbf{y}$ 's which are outside of the domain  $\mathcal{D}$  (in fact,  $\mathbf{y} \in \mathcal{D}^A := \mathcal{D} + (\mathcal{D} - \mathcal{D})$  in the Minkowski sense). On the other hand, by re-using a similar argument as in [Jag13, Lemma 7], we can obtain the same bound (2) for  $C_f^-$ , with the only difference that the Lipschitz constant  $L$  for the gradient function has to be valid on  $\mathcal{D}^A$  instead of just  $\mathcal{D}$ . Finally, the curvature constant for Algorithm 1 is simply the worst-case possibility between the standard FW steps and the away steps:

$$C_f^A := \max\{C_f, C_f^-\}. \quad (16)$$

<sup>8</sup>This can be avoided if the algorithm uses the step-size that minimizes a quadratic upper bound (see the proof for Theorem 7; we can actually use  $\gamma_k := \min\{1, \gamma_{\max}, \frac{g_k}{2C_f}\}$ ); but then one needs to compute an upper bound on  $C_f$  to run the algorithm (which is not always easy). Moreover, this algorithm might have less chance to get the 'best case' behavior by being less adaptive.

**Remark 8.** For all pairs of functions  $f$  and domains  $\mathcal{D}$ , it holds that  $\mu_f^A \leq C_f$  (and  $C_f \leq C_f^A$ ).

*Proof.* Choose  $\mathbf{x}^* := s_f(\mathbf{x})$  for an  $\mathbf{x}$  that is an away corner (i.e.  $\mathbf{x} = \mathbf{v}_f(\mathbf{x})$ ) in (4). Then  $\gamma^A(\mathbf{x}, \mathbf{x}^*) = 1$  and so we have  $\mathbf{y} := \mathbf{x}^* = \mathbf{x} + \gamma(\mathbf{x}^* - \mathbf{x})$  with  $\gamma = 1$  which can also be used in the definition of  $C_f$ . Thus, we have  $\mu_f^A \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq C_f$ .  $\square$

**Theorem' 7.** Suppose that  $f$  has smoothness constant  $C_f^A$  as defined in (16), as well as geometric strong convexity constant  $\mu_f^A$  as defined in (4). Then the error of the iterates of the FW algorithm with away-steps<sup>9</sup> (Algorithm 1) decreases geometrically at each step that is not a drop step (i.e. when  $\gamma_k < \gamma_{\max}$ ), that is

$$h_{k+1} \leq \left(1 - \rho_f^A\right) h_k,$$

where  $\rho_f^A := \frac{\mu_f^A}{4C_f^A}$ . Moreover, the number of drop steps up to iteration  $k$  is bounded by  $k/2$ . This yields the global linear convergence rate of  $h_k \leq h_0 \exp(-\frac{1}{2}\rho_f^A k)$ .

*Proof.* The general idea of the proof is to use the definition of the geometric strong convexity constant to upper bound  $h_k$ , while using the definition of the curvature constant  $C_f^A$  to lower bound the decrease in primal suboptimality  $h_k - h_{k+1}$  for the ‘good steps’ of Algorithm 1. Then we upper bound the number of ‘bad steps’ (the drop steps).

*Upper bounding  $h_k$ .* In the whole proof, we assume that  $\mathbf{x}^{(k)}$  is not already optimal, i.e. that  $h_k > 0$ . If  $h_k = 0$ , then because line-search is used, we will have  $h_{k+1} \leq h_k = 0$  and so the geometric rate of decrease is trivially true in this case.<sup>10</sup> Let  $\mathbf{x}^*$  be an optimum point (which is not necessarily unique). As  $h_k > 0$ , we have that  $\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x}^* - \mathbf{x}^{(k)} \rangle < 0$ . We can thus apply the geometric strong convexity bound (4) at the current iterate  $\mathbf{x} := \mathbf{x}^{(k)}$  using  $\mathbf{x}^*$  as an optimum reference point to get (with  $\bar{\gamma} := \gamma^A(\mathbf{x}^{(k)}, \mathbf{x}^*)$ ):

$$\begin{aligned} \frac{\bar{\gamma}^2}{2} \mu_f^A &\leq f(\mathbf{x}^*) - f(\mathbf{x}^{(k)}) - \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x}^* - \mathbf{x}^{(k)} \rangle \\ &= -h_k - \bar{\gamma} \langle \nabla f(\mathbf{x}^{(k)}), s_f(\mathbf{x}^{(k)}) - \mathbf{v}_f(\mathbf{x}^{(k)}) \rangle \\ &\leq -h_k + \bar{\gamma} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{s}_k - \mathbf{v}_k \rangle \\ &= -h_k + \bar{\gamma} g_k, \end{aligned}$$

where we define  $g_k := \langle -\nabla f(\mathbf{x}^{(k)}), \mathbf{s}_k - \mathbf{v}_k \rangle$  (note that  $h_k \leq g_k$  and so  $g_k$  also gives a primal suboptimality certificate). For the third line, we have used the definition of  $\mathbf{v}_f(\mathbf{x})$  which implies  $\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{v}_f(\mathbf{x}^{(k)}) \rangle \leq \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{v}_k \rangle$ . Therefore  $h_k \leq -\frac{\bar{\gamma}^2}{2} \mu_f^A + \bar{\gamma} g_k$ , which is always upper bounded<sup>11</sup> by  $\frac{g_k^2}{2\mu_f^A}$ :

$$h_k \leq \frac{g_k^2}{2\mu_f^A}. \quad (17)$$

*Lower bounding progress  $h_k - h_{k+1}$ .* A key aspect of the proof is to use the following observation: because of the way the direction  $\mathbf{d}_k$  is chosen in Algorithm 1, we have

$$\langle -\nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k \rangle \geq g_k/2, \quad (18)$$

and thus  $g_k$  characterizes the quality of the direction  $\mathbf{d}_k$ . To see this, note that  $2 \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k \rangle \leq \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k^{\text{FW}} \rangle + \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k^{\text{A}} \rangle = \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k^{\text{FW}} + \mathbf{d}_k^{\text{A}} \rangle = -g_k$ .

We first consider the case  $\gamma_{\max} \geq 1$ . Let  $\mathbf{x}_\gamma := \mathbf{x}^{(k)} + \gamma \mathbf{d}_k$  be the point obtained by moving with step-size  $\gamma$  in direction  $\mathbf{d}_k$ , where  $\mathbf{d}_k$  is the one chosen by Algorithm 1. By using  $\mathbf{s} := \mathbf{x}^{(k)} + \mathbf{d}_k$  (a feasible point as  $\gamma_{\max} \geq 1$ ),  $\mathbf{x} := \mathbf{x}^{(k)}$  and  $\mathbf{y} := \mathbf{x}_\gamma$  in the definition of the curvature constant  $C_f$  (1),

<sup>9</sup>In the algorithm, one can either use line-search or set the step-size as the feasible one that minimizes the quadratic upper bound given by the curvature  $C_f$ , i.e.  $\gamma_k := \min\{1, \gamma_{\max}, \gamma_k^{\text{B}}\}$  where  $\gamma_k^{\text{B}} := \frac{g_k}{2C_f^A}$  and  $g_k := \langle -\nabla f(\mathbf{x}^{(k)}), \mathbf{s}_k - \mathbf{v}_k \rangle$ .

<sup>10</sup>If the fixed schedule step-size is used,  $h_k = 0$  implies that  $g_k = 0$  and so  $\gamma_k = 0$  and thus  $h_{k+1} = h_k$ .

<sup>11</sup>Here we have used the trivial inequality  $0 \leq a^2 - 2ab + b^2$  for the choice of numbers  $a := \frac{g_k}{\mu_f^A}$  and  $b := \bar{\gamma}$ .

and solving for  $f(\mathbf{x}_\gamma)$ , we get  $f(\mathbf{x}_\gamma) \leq f(\mathbf{x}^{(k)}) + \gamma \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k \rangle + \frac{\gamma^2}{2} C_f$ , valid  $\forall \gamma \in [0, 1]$ . As  $\gamma_k$  is obtained by line search and that  $[0, 1] \subseteq [0, \gamma_{\max}]$ , we also have that  $f(\mathbf{x}^{(k+1)}) = f(\mathbf{x}_{\gamma_k}) \leq f(\mathbf{x}_\gamma) \forall \gamma \in [0, 1]$ . Combining these two inequalities, subtracting  $f(\mathbf{x}^*)$  on both sides, and using  $C_f \leq C_f^A$  to simplify the possibilities yields  $h_{k+1} \leq h_k + \gamma \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k \rangle + \frac{\gamma^2}{2} C_f^A$ .

Using the crucial gap inequality (18), we get  $h_{k+1} \leq h_k - \gamma \frac{g_k}{2} + \frac{\gamma^2}{2} C_f^A$ , and so:

$$h_k - h_{k+1} \geq \gamma \frac{g_k}{2} - \frac{\gamma^2}{2} C_f^A \quad \forall \gamma \in [0, 1]. \quad (19)$$

We can minimize the bound (19) on the right hand side by letting  $\gamma = \gamma_k^B := \frac{g_k}{2C_f^A}$  – supposing that  $\gamma_k^B \leq 1$ , we then get  $h_k - h_{k+1} \geq \frac{g_k^2}{8C_f^A}$  (we cover the case  $\gamma_k^B > 1$  later). By combining this inequality with the one from geometric strong convexity (17), we get

$$\frac{h_k - h_{k+1}}{h_k} \geq \frac{\mu_f^A}{4C_f^A} \quad (20)$$

implying that we have a geometric rate of decrease  $h_{k+1} \leq \left(1 - \frac{\mu_f^A}{4C_f^A}\right) h_k$  (this is a ‘good step’).

*Boundary cases.* We now consider the case  $\gamma_k^B > 1$  (with  $\gamma_{\max} \geq 1$  still). The condition  $\gamma_k^B > 1$  then translates to  $g_k \geq 2C_f^A$ , which we can use in (19) with  $\gamma = 1$  to get  $h_k - h_{k+1} \geq \frac{g_k}{2} - \frac{g_k}{4} = \frac{g_k}{4}$ . Combining this inequality with  $h_k \leq g_k$  gives the geometric decrease  $h_{k+1} \leq \left(1 - \frac{1}{4}\right) h_k$  (also a ‘good step’).  $\rho_f^A$  is obtained by considering the worst-case of the constants obtained from  $\gamma_k^B > 1$  and  $\gamma_k^B \leq 1$ . (Note that always  $\mu_f^A \leq C_f^A$  by definition, as discussed in Remark 8).

Finally, we are left with the case that  $\gamma_{\max} < 1$ . This is thus an away step and so  $\mathbf{d}_k = \mathbf{d}_k^A = \mathbf{x}^{(k)} - \mathbf{v}_k$ . Here, we use the away version  $C_f^-$  of the definition for  $C_f^A$ : by letting  $\mathbf{s} := \mathbf{v}_k$ ,  $\mathbf{x} := \mathbf{x}^{(k)}$  and  $\mathbf{y} := \mathbf{x}_\gamma$  in (15), we also get the bound  $f(\mathbf{x}_\gamma) \leq f(\mathbf{x}^{(k)}) + \gamma \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d}_k \rangle + \frac{\gamma^2}{2} C_f^A$ , valid  $\forall \gamma \in [0, 1]$  (but note here that the points  $\mathbf{x}_\gamma$  are not feasible for  $\gamma > \gamma_{\max}$  – the bound considers some points outside of  $\mathcal{D}$ ). We now have two options: either  $\gamma_k = \gamma_{\max}$  (a drop step) or  $\gamma_k < \gamma_{\max}$ . In the case  $\gamma_k < \gamma_{\max}$  (the line-search yields a solution in the interior of  $[0, \gamma_{\max}]$ ), then because  $f(\mathbf{x}_\gamma)$  is convex in  $\gamma$ , we know that  $\min_{\gamma \in [0, \gamma_{\max}]} f(\mathbf{x}_\gamma) = \min_{\gamma \geq 0} f(\mathbf{x}_\gamma)$  and thus  $\min_{\gamma \in [0, \gamma_{\max}]} f(\mathbf{x}_\gamma) = f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}_\gamma) \forall \gamma \in [0, 1]$ . We can then re-use the same argument above equation (19) to get the inequality (19), and again considering both the case  $\gamma_k^B \leq 1$  (which yields inequality (20)) and the case  $\gamma_k^B > 1$  (which yields  $(1 - \frac{1}{4})$  as the geometric rate constant), we get a ‘good step’ with  $1 - \rho_f^A$  as the worst-case geometric rate constant.

Finally, we can easily bound the number of drop steps possible up to iteration  $k$  with the following argument (the drop steps are the ‘bad steps’ for which we cannot show good progress). Let  $A_k$  be the number of steps that added a vertex in the expansion (only standard FW steps can do this) and let  $D_k$  be the number of drop steps. We have that  $|\mathcal{S}^{(k)}| = |\mathcal{S}^{(0)}| + A_k - D_k$ . Moreover, we have that  $A_k + D_k \leq k$ . We thus have  $1 \leq |\mathcal{S}^{(k)}| \leq |\mathcal{S}^{(0)}| + k - 2D_k$ , implying that  $D_k \leq \frac{1}{2}(|\mathcal{S}^{(0)}| - 1 + k) = \frac{k}{2}$ , as stated in the theorem.  $\square$