

On the Global Linear Convergence of Frank-Wolfe Optimization Variants

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

Martin Jaggi
Dept. of Computer Science
ETH Zürich, Switzerland

Abstract

The Frank-Wolfe (FW) optimization algorithm has lately re-gained popularity thanks in particular to its ability to nicely handle the structured constraints appearing in machine learning applications. However, its convergence rate is known to be slow (sublinear) when the solution lies at the boundary. A simple less-known fix is to add the possibility to take ‘away steps’ during optimization, an operation that importantly *does not* require a feasibility oracle. In this paper, we highlight and clarify several variants of the Frank-Wolfe optimization algorithm that have been successfully applied in practice: away-steps FW, pairwise FW, fully-corrective FW and Wolfe’s minimum norm point algorithm, and prove for the first time that they all enjoy global linear convergence, under a weaker condition than strong convexity of the objective. The constant in the convergence rate has an elegant interpretation as the product of the (classical) condition number of the function with a novel geometric quantity that plays the role of a ‘condition number’ of the constraint set. We provide pointers to where these algorithms have made a difference in practice, in particular with the flow polytope, the marginal polytope and the base polytope for submodular optimization.

The Frank-Wolfe algorithm [9] (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 t iterations will decrease geometrically with $O((1 - \rho)^t)$ instead of the usual $O(1/t)$ 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 clarify several variants of the Frank-Wolfe algorithm and show that they all converge linearly for any strongly convex function optimized over a polytope domain, with a constant bounded away from zero that only depends on the geometry of the polytope. 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 [34, 12, 5], and the newer work of [28], as well as the line of work of [1, 19, 26] which rely on Robinson’s condition [30]. 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. The constant also naturally separates as the product of the condition number of the function with a novel notion of condition number of a polytope, which might have applications in complexity theory.

Related Work. For the classical Frank-Wolfe algorithm, [5] showed a linear rate for the special case of quadratic objectives when the optimum is in the strict interior of the domain, a result already subsumed by the more general [12]. The early work of [23] showed linear convergence for *strongly*

convex constraint sets, under the strong requirement that the gradient norm is not too small (see [11] for a discussion). The away-steps variant of the Frank-Wolfe algorithm, that can also remove weight from ‘bad’ atoms in the current active set, was proposed in [34], and later also analyzed in [12]. The precise method is stated below in Algorithm 1. [12] showed a (local) linear convergence rate on polytopes, but the constant unfortunately depends on the distance between the solution and its relative boundary, a quantity that can be arbitrarily small. More recently, [1, 19, 26] have obtained linear convergence results in the case that the optimum solution satisfies Robinson’s condition [30]. In a different recent line of work, [10, 22] have studied a variation of FW that repeatedly moves mass from the worst vertices to the standard FW vertex until a specific condition is satisfied, yielding a linear rate on strongly convex functions. 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.

Setup. We consider general constrained convex optimization problems of the form:

$$\min_{\mathbf{x} \in \mathcal{M}} f(\mathbf{x}), \quad \mathcal{M} = \text{conv}(\mathcal{A}), \quad \text{with only access to: } \text{LMO}_{\mathcal{A}}(\mathbf{r}) \in \arg \min_{\mathbf{x} \in \mathcal{A}} \langle \mathbf{r}, \mathbf{x} \rangle, \quad (1)$$

where $\mathcal{A} \subseteq \mathbb{R}^d$ is a *finite* set of vectors that we call *atoms*.¹ We assume that the function f is μ -strongly convex with L -Lipschitz continuous gradient over \mathcal{M} . We also consider weaker conditions than strong convexity for f in Section 4. As \mathcal{A} is finite, \mathcal{M} is a (convex and bounded) polytope. The methods that we consider in this paper only require access to a *linear minimization oracle* $\text{LMO}_{\mathcal{A}}(\cdot)$ associated with the domain \mathcal{M} through a generating set of atoms \mathcal{A} . This oracle is defined as to return a minimizer of a linear subproblem over $\mathcal{M} = \text{conv}(\mathcal{A})$, for any given direction $\mathbf{r} \in \mathbb{R}^d$.²

Examples. Optimization problems of the form (1) appear widely in machine learning and signal processing applications. The set of atoms \mathcal{A} can represent combinatorial objects of arbitrary type. Efficient linear minimization oracles often exist in the form of dynamic programs or other combinatorial optimization approaches. As an example from tracking in computer vision, \mathcal{A} could be the set of integer flows on a graph [16, 7], where $\text{LMO}_{\mathcal{A}}$ can be efficiently implemented by a minimum cost network flow algorithm. In this case, \mathcal{M} can also be described with a polynomial number of linear inequalities. But in other examples, \mathcal{M} might not have a polynomial description in terms of linear inequalities, and testing membership in \mathcal{M} might be much more expensive than running the linear oracle. This is the case when optimizing over the *base polytope*, an object appearing in submodular function optimization [3]. There, the $\text{LMO}_{\mathcal{A}}$ oracle is a simple greedy algorithm. Another example is when \mathcal{A} represents the possible consistent value assignments on cliques of a Markov random field (MRF); \mathcal{M} is the *marginal polytope* [32], where testing membership is NP-hard in general, though efficient linear oracles exist for some special cases [17]. Optimization over the marginal polytope appears for example in structured SVM learning [21] and variational inference [18].

The Original Frank-Wolfe Algorithm. The Frank-Wolfe (FW) optimization algorithm [9], also known as *conditional gradient* [23], is particularly suited for the setup (1) where \mathcal{M} is only accessed through the linear minimization oracle. It works as follows: At a current iterate $\mathbf{x}^{(t)}$, the algorithm finds a feasible search atom \mathbf{s}_t to move towards by minimizing the linearization of the objective function f over \mathcal{M} (line 3 in Algorithm 1) – this is where the linear minimization oracle $\text{LMO}_{\mathcal{A}}$ is used. The next iterate $\mathbf{x}^{(t+1)}$ is then obtained by doing a *line-search* on f between $\mathbf{x}^{(t)}$ and \mathbf{s}_t (line 11 in Algorithm 1). One reason for the recent increased popularity of Frank-Wolfe-type algorithms is the sparsity of their iterates: in iteration t of the algorithm, the iterate can be represented as a sparse convex combination of at most $t + 1$ atoms $\mathcal{S}^{(t)} \subseteq \mathcal{A}$ of the domain \mathcal{M} , which we write as $\mathbf{x}^{(t)} = \sum_{\mathbf{v} \in \mathcal{S}^{(t)}} \alpha_{\mathbf{v}}^{(t)} \mathbf{v}$. We write $\mathcal{S}^{(t)}$ for the *active set*, containing the previously discovered search atoms \mathbf{s}_r for $r < t$ that have non-zero *weight* $\alpha_{\mathbf{s}_r}^{(t)} > 0$ in the expansion (potentially also including the starting point $\mathbf{x}^{(0)}$). While tracking the active set $\mathcal{S}^{(t)}$ is not necessary for the original FW algorithm, the improved variants of FW that we discuss will require that $\mathcal{S}^{(t)}$ is maintained.

Zig-Zagging Phenomenon. When the optimal solution lies at the boundary of \mathcal{M} , the convergence rate of the iterates is slow, i.e. sublinear: $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq O(1/t)$, for \mathbf{x}^* being an optimal solution [9, 6, 8, 15]. This is because the iterates of the classical FW algorithm start to zig-zag

¹The atoms *do not* have to be extreme points (vertices) of \mathcal{M} .

²All our convergence results can be carefully extended to approximate linear minimization oracles with multiplicative approximation guarantees; we state them for exact oracles in this paper for simplicity.

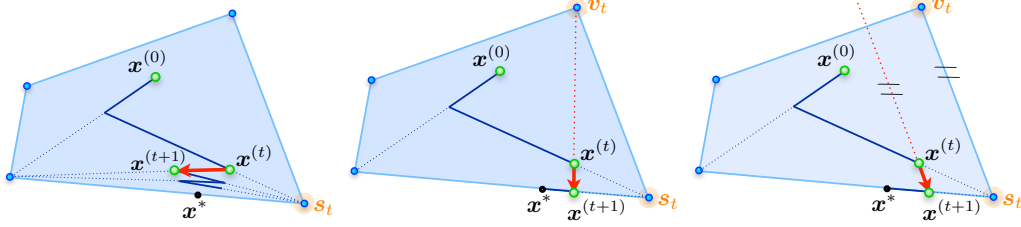


Figure 1: (left) The FW algorithm zig-zags when the solution x^* lies on the boundary. (middle) Adding the possibility of an *away step* attenuates this problem. (right) As an alternative, a pairwise FW step.

between the vertices defining the face containing the solution x^* (see left of Figure 1). In fact, the $1/t$ rate is tight for a large class of functions: Canon and Cullum [6], Wolfe [34] showed (roughly) that $f(x^{(t)}) - f(x^*) \geq \Omega(1/t^{1+\delta})$ for any $\delta > 0$ when x^* lies on a face of \mathcal{M} with some additional regularity assumptions. Note that this lower bound is different than the $\Omega(1/t)$ one presented in [15, Lemma 3] which holds for all one-atom-per-step algorithms but assumes high dimensionality $d \geq t$.

1 Improved Variants of the Frank-Wolfe Algorithm

Algorithm 1 Away-steps Frank-Wolfe algorithm: $\text{AFW}(x^{(0)}, \mathcal{A}, \epsilon)$

```

1: Let  $x^{(0)} \in \mathcal{A}$ , and  $\mathcal{S}^{(0)} := \{x^{(0)}\}$  (so that  $\alpha_v^{(0)} = 1$  for  $v = x^{(0)}$  and 0 otherwise)
2: for  $t = 0 \dots T$  do
3:   Let  $s_t := \text{LMO}_{\mathcal{A}}(\nabla f(x^{(t)}))$  and  $d_t^{\text{FW}} := s_t - x^{(t)}$  (the FW direction)
4:   Let  $v_t \in \arg \max_{v \in \mathcal{S}^{(t)}} \langle \nabla f(x^{(t)}), v \rangle$  and  $d_t^{\text{A}} := x^{(t)} - v_t$  (the away direction)
5:   if  $g_t^{\text{FW}} := \langle -\nabla f(x^{(t)}), d_t^{\text{FW}} \rangle \leq \epsilon$  then return  $x^{(t)}$  (FW gap is small enough, so return)
6:   if  $\langle -\nabla f(x^{(t)}), d_t^{\text{FW}} \rangle \geq \langle -\nabla f(x^{(t)}), d_t^{\text{A}} \rangle$  then
7:      $d_t := d_t^{\text{FW}}$ , and  $\gamma_{\max} := 1$  (choose the FW direction)
8:   else
9:      $d_t := d_t^{\text{A}}$ , and  $\gamma_{\max} := \alpha_{v_t} / (1 - \alpha_{v_t})$  (choose away direction; maximum feasible step-size)
10:  end if
11:  Line-search:  $\gamma_t \in \arg \min_{\gamma \in [0, \gamma_{\max}]} f(x^{(t)} + \gamma d_t)$ 
12:  Update  $x^{(t+1)} := x^{(t)} + \gamma_t d_t$  (and accordingly for the weights  $\alpha^{(t+1)}$ , see text)
13:  Update  $\mathcal{S}^{(t+1)} := \{v \in \mathcal{A} \text{ s.t. } \alpha_v^{(t+1)} > 0\}$ 
14: end for

```

Algorithm 2 Pairwise Frank-Wolfe algorithm: $\text{PFW}(x^{(0)}, \mathcal{A}, \epsilon)$

```

1: ... as in Algorithm 1, except replacing lines 6 to 10 by:  $d_t = d_t^{\text{PFW}} := s_t - v_t$ , and  $\gamma_{\max} := \alpha_{v_t}$ .

```

Away-Steps Frank-Wolfe. To address the zig-zagging problem of FW, Wolfe [34] proposed to add the possibility to move away from an active atom in $\mathcal{S}^{(t)}$ (see middle of Figure 1); this simple modification is sufficient to make the algorithm linearly convergent for strongly convex functions. We describe the away-steps variant of Frank-Wolfe in Algorithm 1.³ The away direction d_t^{A} is defined in line 4 by finding the atom v_t in $\mathcal{S}^{(t)}$ that maximizes the potential of descent given by $g_t^{\text{A}} := \langle -\nabla f(x^{(t)}), x^{(t)} - v_t \rangle$. Note that this search is over the (typically small) active set $\mathcal{S}^{(t)}$, and is fundamentally easier than the linear oracle $\text{LMO}_{\mathcal{A}}$. The maximum step-size γ_{\max} as defined on line 9 ensures that the new iterate $x^{(t)} + \gamma d_t^{\text{A}}$ stays in \mathcal{M} . In fact, this guarantees that the convex representation is maintained, and we stay inside $\text{conv}(\mathcal{S}^{(t)}) \subseteq \mathcal{M}$. When \mathcal{M} is a simplex, then the barycentric coordinates are unique and $x^{(t)} + \gamma_{\max} d_t^{\text{A}}$ truly lies on the boundary of \mathcal{M} . On the other hand, if $|\mathcal{A}| > \dim(\mathcal{M}) + 1$ (e.g. for the cube), then it could hypothetically be possible to have a step-size bigger than γ_{\max} which is still feasible. Computing the true maximum feasible step-size would require the ability to know when we cross the boundary of \mathcal{M} along a specific line, which is not possible for general \mathcal{M} . Using the conservative maximum step-size of line 9 ensures that we

³The original algorithm presented in [34] was not convergent; this was corrected by Guélat and Marcotte [12], assuming a tractable representation of \mathcal{M} with linear inequalities and called it the modified Frank-Wolfe (MFW) algorithm. Our description in Algorithm 1 extends it to the more general setup of (1).

do not need this more powerful oracle. This is why Algorithm 1 requires to maintain $\mathcal{S}^{(t)}$ (unlike standard FW). Finally, as in classical FW, the FW gap g_t^{FW} is an upper bound on the unknown suboptimality, and can be used as a stopping criterion:

$$g_t^{\text{FW}} := \langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t^{\text{FW}} \rangle \geq \langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{x}^* - \mathbf{x}^{(t)} \rangle \geq f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \quad (\text{by convexity}).$$

If $\gamma_t = \gamma_{\max}$, then we call this step a *drop step*, as it fully removes the atom \mathbf{v}_t from the currently active set of atoms $\mathcal{S}^{(t)}$ (by settings its weight to zero). The weight updates for lines 12 and 13 are of the following form: For a FW step, we have $\mathcal{S}^{(t+1)} = \{\mathbf{s}_t\}$ if $\gamma_t = 1$; otherwise $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} \cup \{\mathbf{s}_t\}$. Also, we have $\alpha_{\mathbf{s}_t}^{(t+1)} := (1 - \gamma_t)\alpha_{\mathbf{s}_t}^{(t)} + \gamma_t$ and $\alpha_{\mathbf{v}}^{(t+1)} := (1 - \gamma_t)\alpha_{\mathbf{v}}^{(t)}$ for $\mathbf{v} \in \mathcal{S}^{(t)} \setminus \{\mathbf{s}_t\}$. For an away step, we have $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} \setminus \{\mathbf{v}_t\}$ if $\gamma_t = \gamma_{\max}$ (a *drop step*); otherwise $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)}$. Also, we have $\alpha_{\mathbf{v}_t}^{(t+1)} := (1 + \gamma_t)\alpha_{\mathbf{v}_t}^{(t)} - \gamma_t$ and $\alpha_{\mathbf{v}}^{(t+1)} := (1 + \gamma_t)\alpha_{\mathbf{v}}^{(t)}$ for $\mathbf{v} \in \mathcal{S}^{(t)} \setminus \{\mathbf{v}_t\}$.

Pairwise Frank-Wolfe. The next variant that we present is inspired by an early algorithm by Mitchell et al. [25], called the MDM algorithm, originally invented for the polytope distance problem. Here the idea is to only move weight mass between two atoms in each step. More precisely, the generalized method as presented in Algorithm 2 moves weight from the away atom \mathbf{v}_t to the FW atom \mathbf{s}_t , and keeps all other α weights un-changed. We call such a swap of mass between the two atoms a *pairwise FW step*, i.e. $\alpha_{\mathbf{v}_t}^{(t+1)} = \alpha_{\mathbf{v}_t}^{(t)} - \gamma$ and $\alpha_{\mathbf{s}_t}^{(t+1)} = \alpha_{\mathbf{s}_t}^{(t)} + \gamma$ for some step-size $\gamma \leq \gamma_{\max} := \alpha_{\mathbf{v}_t}^{(t)}$. In contrast, classical FW shrinks all active weights at every iteration.

The pairwise FW direction will also be central to our proof technique to provide the first global linear convergence rate for away-steps FW, as well as the fully-corrective variant and Wolfe’s min-norm-point algorithm.

As we will see in Section 2.2, the rate guarantee for the pairwise FW variant is more loose than for the other variants, because we cannot provide a satisfactory bound on the number of the problematic *swap steps* (defined just before Theorem 1). Nevertheless, the algorithm seems to perform quite well in practice, often outperforming away-steps FW, especially in the important case of sparse solutions, that is if the optimal solution \mathbf{x}^* lies on a low-dimensional face of \mathcal{M} (and thus one wants to keep the active set $\mathcal{S}^{(t)}$ small). The pairwise FW step is arguably more efficient at pruning the coordinates in $\mathcal{S}^{(t)}$. In contrast to the away step which moves the mass back *uniformly* onto all other active elements $\mathcal{S}^{(t)}$ (and might require more corrections later), the pairwise FW step only moves the mass onto the (good) FW atom \mathbf{s}_t . A slightly different version than Algorithm 2 was also proposed by Ľanculef et al. [26], though their convergence proofs were incomplete (see Appendix A.3). The algorithm is related to classical working set algorithms, such as the SMO algorithm used to train SVMs [29]. We refer to [26] for an empirical comparison for SVMs, as well as their Section 5 for more related work. See also Appendix A.3 for a link between pairwise FW and [10].

Fully-Corrective Frank-Wolfe, and Wolfe’s Min-Norm Point Algorithm. When the linear oracle is expensive, it might be worthwhile to do more work to optimize over the active set $\mathcal{S}^{(t)}$ in between each call to the linear oracle, rather than just performing an away or pairwise step. We give in Algorithm 3 the fully-corrective Frank-Wolfe (FCFW) variant, that maintains a correction polytope defined by a set of atoms $\mathcal{A}^{(t)}$ (potentially larger than the active set $\mathcal{S}^{(t)}$). Rather than obtaining the next iterate by line-search, $\mathbf{x}^{(t+1)}$ is obtained by re-optimizing f over $\text{conv}(\mathcal{A}^{(t)})$. Depending on how the correction is implemented, and how the correction atoms $\mathcal{A}^{(t)}$ are maintained, several variants can be obtained. These variants are known under many names, such as the extended FW method by Holloway [14] or the simplicial decomposition method [31, 13]. Wolfe’s min-norm point (MNP) algorithm [35] for polytope distance problems is often confused with FCFW for quadratic objectives. The major difference is that standard FCFW optimizes f over $\text{conv}(\mathcal{A}^{(t)})$, whereas MNP implements the correction as a sequence of affine projections that potentially yield a different update, but can be computed more efficiently in several practical applications [35]. We describe precisely in Appendix A.1 a generalization of the MNP algorithm as a specific case of the correction subroutine from step 7 of the generic Algorithm 3.

The original convergence analysis of the FCFW algorithm [14] (and also MNP algorithm [35]) only showed that they were finitely convergent, with a bound on the number of iterations in terms of the cardinality of \mathcal{A} (unfortunately an exponential number in general). Holloway [14] also argued that FCFW had an asymptotic linear convergence based on the flawed argument of Wolfe [34]. As far as we know, our work is the first to provide global linear convergence rates for FCFW and MNP for

Algorithm 3 Fully-corrective Frank-Wolfe with approximate correction: **FCFW**($\mathbf{x}^{(0)}, \mathcal{A}, \epsilon$)

- 1: **Input:** Set of atoms \mathcal{A} , active set $\mathcal{S}^{(0)}$, starting point $\mathbf{x}^{(0)} = \sum_{\mathbf{v} \in \mathcal{S}^{(0)}} \alpha_{\mathbf{v}}^{(0)} \mathbf{v}$, stopping criterion ϵ .
 - 2: Let $\mathcal{A}^{(0)} := \mathcal{S}^{(0)}$ (optionally, a bigger $\mathcal{A}^{(0)}$ could be passed as argument for a warm start)
 - 3: **for** $t = 0 \dots T$ **do**
 - 4: Let $\mathbf{s}_t := \text{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}^{(t)}))$ *(the FW atom)*
 - 5: Let $\mathbf{d}_t^{\text{FW}} := \mathbf{s}_t - \mathbf{x}^{(t)}$ and $g_t^{\text{FW}} = \langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t^{\text{FW}} \rangle$ *(FW gap)*
 - 6: **if** $g_t^{\text{FW}} \leq \epsilon$ **then return** $\mathbf{x}^{(t)}$
 - 7: $(\mathbf{x}^{(t+1)}, \mathcal{A}^{(t+1)}) := \text{Correction}(\mathbf{x}^{(t)}, \mathcal{A}^{(t)}, \mathbf{s}_t, \epsilon)$ *(approximate correction step)*
 - 8: **end for**
-

Algorithm 4 Approximate correction: **Correction**($\mathbf{x}^{(t)}, \mathcal{A}^{(t)}, \mathbf{s}_t, \epsilon$)

- 1: Return $(\mathbf{x}^{(t+1)}, \mathcal{A}^{(t+1)})$ with the following properties:
 - 2: $\mathcal{S}^{(t+1)}$ is the active set for $\mathbf{x}^{(t+1)}$ and $\mathcal{A}^{(t+1)} \supseteq \mathcal{S}^{(t+1)}$.
 - 3: $f(\mathbf{x}^{(t+1)}) \leq \min_{\gamma \in [0,1]} f(\mathbf{x}^{(t)} + \gamma(\mathbf{s}_t - \mathbf{x}^{(t)}))$ *(make at least as much progress as a FW step)*
 - 4: $g_{t+1}^{\mathcal{A}} := \max_{\mathbf{v} \in \mathcal{S}^{(t+1)}} \langle -\nabla f(\mathbf{x}^{(t+1)}), \mathbf{x}^{(t+1)} - \mathbf{v} \rangle \leq \epsilon$ *(the away gap is small enough)*
-

general strongly convex functions. Moreover, the proof of convergence for FCFW does not require an exact solution to the correction step; instead, we show that the weaker properties stated for the approximate correction procedure in Algorithm 4 are sufficient for a global linear convergence rate (this correction could be implemented using away-steps FW, as done for example in [18]).

2 Global Linear Convergence Analysis

2.1 Intuition for the Convergence Proofs

We first give the general intuition for the linear convergence proof of the different FW variants, starting from the work of Guélat and Marcotte [12]. We assume that the objective function f is smooth over a compact set \mathcal{M} , i.e. its gradient is Lipschitz continuous with constant L . Also let $M := \text{diam}(\mathcal{M})$. Let \mathbf{d}_t be the direction in which the line-search is executed by the algorithm (Line 11 in Algorithm 1). By the standard descent lemma [see e.g. (1.2.5) in 27], we have:

$$f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}^{(t)} + \gamma \mathbf{d}_t) \leq f(\mathbf{x}^{(t)}) + \gamma \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t \rangle + \frac{\gamma^2}{2} L \|\mathbf{d}_t\|^2 \quad \forall \gamma \in [0, \gamma_{\max}]. \quad (2)$$

We let $\mathbf{r}_t := -\nabla f(\mathbf{x}^{(t)})$ and let $h_t := f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)$ be the suboptimality error. Supposing for now that $\gamma_{\max} \geq \gamma_t^* := \langle \mathbf{r}_t, \mathbf{d}_t \rangle / (L \|\mathbf{d}_t\|^2)$. We can set $\gamma = \gamma_t^*$ to minimize the RHS of (2), subtract $f(\mathbf{x}^*)$ on both sides, and re-organize to get a lower bound on the progress:

$$h_t - h_{t+1} \geq \frac{\langle \mathbf{r}_t, \mathbf{d}_t \rangle^2}{2L \|\mathbf{d}_t\|^2} = \frac{1}{2L} \langle \mathbf{r}_t, \hat{\mathbf{d}}_t \rangle^2, \quad (3)$$

where we use the ‘hat’ notation to denote normalized vectors: $\hat{\mathbf{d}}_t := \mathbf{d}_t / \|\mathbf{d}_t\|$. Let $\mathbf{e}_t := \mathbf{x}^* - \mathbf{x}^{(t)}$ be the error vector. By μ -strong convexity of f , we have:

$$f(\mathbf{x}^{(t)} + \gamma \mathbf{e}_t) \geq f(\mathbf{x}^{(t)}) + \gamma \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}_t \rangle + \frac{\gamma^2}{2} \mu \|\mathbf{e}_t\|^2 \quad \forall \gamma \in [0, 1]. \quad (4)$$

The RHS is lower bounded by its minimum as a function of γ (unconstrained), achieved using $\gamma := \langle \mathbf{r}_t, \mathbf{e}_t \rangle / (\mu \|\mathbf{e}_t\|^2)$. We are then free to use any value of γ on the LHS and maintain a valid bound. In particular, we use $\gamma = 1$ to obtain $f(\mathbf{x}^*)$. Again re-arranging, we get:

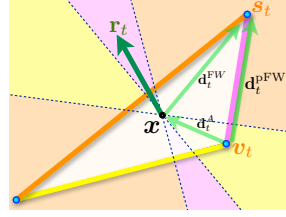
$$h_t \leq \frac{\langle \mathbf{r}_t, \hat{\mathbf{e}}_t \rangle^2}{2\mu}, \quad \text{and combining with (3), we obtain:} \quad h_t - h_{t+1} \geq \frac{\mu}{L} \frac{\langle \mathbf{r}_t, \hat{\mathbf{d}}_t \rangle^2}{\langle \mathbf{r}_t, \hat{\mathbf{e}}_t \rangle^2} h_t. \quad (5)$$

The inequality (5) is fairly general and valid for any line-search method in direction \mathbf{d}_t . To get a linear convergence rate, we need to lower bound (by a positive constant) the term in front of h_t on the RHS, which depends on the angle between the update direction \mathbf{d}_t and the negative gradient \mathbf{r}_t . If we assume that the solution \mathbf{x}^* lies in the relative interior of \mathcal{M} with a distance of at least $\delta > 0$ from the boundary, then $\langle \mathbf{r}_t, \mathbf{d}_t \rangle \geq \delta \|\mathbf{r}_t\|$ for the FW direction \mathbf{d}_t^{FW} , and by combining with $\|\mathbf{d}_t\| \leq M$, we get a linear rate with constant $1 - \frac{\mu}{L} (\frac{\delta}{M})^2$ (this was the result from [12]). On the other hand, if \mathbf{x}^* lies on the boundary, then $\langle \hat{\mathbf{r}}_t, \hat{\mathbf{d}}_t \rangle$ gets arbitrary close to zero for standard FW (the zig-zagging phenomenon) and the convergence is sublinear.

Proof Sketch for AFW. The key insight to prove the **global linear convergence** for AFW is to relate $\langle \mathbf{r}_t, \mathbf{d}_t \rangle$ with the *pairwise FW* direction $\mathbf{d}_t^{\text{PFW}} := \mathbf{s}_t - \mathbf{v}_t$. By the way the direction \mathbf{d}_t is chosen on lines 6 to 10 of Algorithm 1, we have:

$$2\langle \mathbf{r}_t, \mathbf{d}_t \rangle \geq \langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} \rangle + \langle \mathbf{r}_t, \mathbf{d}_t^{\text{A}} \rangle = \langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} + \mathbf{d}_t^{\text{A}} \rangle = \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle. \quad (6)$$

We thus have $\langle \mathbf{r}_t, \mathbf{d}_t \rangle \geq \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle / 2$. Now the crucial property of the pairwise FW direction is that for any potential negative gradient direction \mathbf{r}_t , the worst case inner product $\langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle$ can be lower bounded away from zero by a quantity depending only on the geometry of \mathcal{M} (unless we are at the optimum). We call this quantity the *pyramidal width* of \mathcal{A} . The figure on the right shows the six possible pairwise FW directions $\mathbf{d}_t^{\text{PFW}}$ for a triangle domain, depending on which colored area the \mathbf{r}_t direction falls into. We will see that the pyramidal width is related to the smallest width of pyramids that we can construct from \mathcal{A} in a specific way related to the choice of the away and towards atoms \mathbf{v}_t and \mathbf{s}_t . See (9) and our main Theorem 3 in Section 3.



This gives the main argument for the linear convergence of AFW for steps where $\gamma_t^* \leq \gamma_{\max}$. When γ_{\max} is too small, AFW will perform a *drop step*, as the line-search will truncate the step-size to $\gamma_t = \gamma_{\max}$. We cannot guarantee sufficient progress in this case, but the drop step decreases the active set size by one, and thus they cannot happen too often (not more than half the time). These are the main elements for the global linear convergence proof for AFW. The rest is to carefully consider various boundary cases. We can re-use the same techniques to prove the convergence for pairwise FW, though unfortunately the latter also has the possibility of problematic *swap steps*. While their number can be bounded, so far we only found the extremely loose bound quoted in Theorem 1.

Proof Sketch for FCFW. For FCFW, by line 4 of the correction Algorithm 4, the away gap satisfies $g_t^{\text{A}} \leq \epsilon$ at the beginning of a new iteration. Supposing that the algorithm does not exit at line 6 of Algorithm 3, we have $g_t^{\text{FW}} > \epsilon$ and therefore $2\langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} \rangle \geq \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle$ using a similar argument as in (6). Finally, by line 3 of Algorithm 4, the correction is guaranteed to make at least as much progress as a line-search in direction \mathbf{d}_t^{FW} , and so the progress bound (5) applies also to FCFW.

2.2 Convergence Results

We now give the global linear convergence rates for the four variants of the FW algorithm: away-steps FW (AFW Alg. 1); pairwise FW (PFW Alg. 2); fully-corrective FW (FCFW Alg. 3 with approximate correction Alg. 4); and Wolfe's min-norm point algorithm (Alg. 3 with MNP-correction as Alg. 5 in Appendix A.1). For the AFW, MNP and PFW algorithms, we call a *drop step* when the active set shrinks $|S^{(t+1)}| < |S^{(t)}|$. For the PFW algorithm, we also have the possibility of a *swap step* where $\gamma_t = \gamma_{\max}$ but $|S^{(t+1)}| = |S^{(t)}|$ (i.e. the mass was fully swapped from the away atom to the FW atom). A nice property of FCFW is that it does not have any drop step (it executes both FW steps and away steps simultaneously while guaranteeing enough progress at every iteration).

Theorem 1. Suppose that f has L -Lipschitz gradient⁴ and is μ -strongly convex over $\mathcal{M} = \text{conv}(\mathcal{A})$. Let $M = \text{diam}(\mathcal{M})$ and $\delta = \text{PWidth}(\mathcal{A})$ as defined by (9). Then the suboptimality h_t of the iterates of all the four variants of the FW algorithm decreases geometrically at each step that is not a drop step nor a swap step (i.e. when $\gamma_t < \gamma_{\max}$, called a 'good step'), that is

$$h_{t+1} \leq (1 - \rho) h_t, \quad \text{where } \rho := \frac{\mu}{4L} \left(\frac{\delta}{M} \right)^2.$$

Let $k(t)$ be the number of 'good steps' up to iteration t . We have $k(t) = t$ for FCFW; $k(t) \geq t/2$ for MNP and AFW; and $k(t) \geq t/(3|\mathcal{A}| + 1)$ for PFW (because of the swap steps). This yields a global linear convergence rate of $h_t \leq h_0 \exp(-\rho k(t))$ for all variants. If $\mu = 0$ (general convex), then $h_t = O(1/k(t))$ instead. See Theorem 8 in Appendix D for an affine invariant version and proof.

Note that to our knowledge, none of the existing linear convergence results showed that the duality gap was also linearly convergent. The result for the gap follows directly from the simple manipulation of (2); putting the FW gap to the LHS and optimizing the RHS for $\gamma \in [0, 1]$.

Theorem 2. Suppose that f has L -Lipschitz gradient over \mathcal{M} with $M := \text{diam}(\mathcal{M})$. Then the FW gap g_t^{FW} for any algorithm is upper bounded by the primal error h_t as follows:

$$g_t^{\text{FW}} \leq h_t + LM^2/2 \text{ when } h_t > LM^2/2, \quad g_t^{\text{FW}} \leq M\sqrt{2h_tL} \text{ otherwise.} \quad (7)$$

⁴For AFW and PFW, we actually require that ∇f is L -Lipschitz over the larger domain $\mathcal{M} + \mathcal{M} - \mathcal{M}$.

3 Pyramidal Width

We now describe the claimed lower bound on the angle between the negative gradient and the pairwise FW direction, which depends only on the geometric properties of \mathcal{M} . According to our argument about the progress bound (5) and the PFW gap (6), our goal is to find a lower bound on $\langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle / \langle \mathbf{r}_t, \hat{\mathbf{e}}_t \rangle$. First note that $\langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle = \langle \mathbf{r}_t, \mathbf{s}_t - \mathbf{v}_t \rangle = \max_{\mathbf{s} \in \mathcal{M}, \mathbf{v} \in \mathcal{S}^{(t)}} \langle \mathbf{r}_t, \mathbf{s} - \mathbf{v} \rangle$ where $\mathcal{S}^{(t)}$ is a pos-

sible active set for $\mathbf{x}^{(t)}$. This looks like the *directional width* of a pyramid with base $\mathcal{S}^{(t)}$ and summit \mathbf{s}_t . To be conservative, we consider the worst case possible active set for $\mathbf{x}^{(t)}$; this is what we will call the *pyramidal directional width* $PdirW(\mathcal{A}, \mathbf{r}_t, \mathbf{x}^{(t)})$. We start with the following definitions.

Directional Width. The directional width of a set \mathcal{A} with respect to a direction \mathbf{r} is defined as $dirW(\mathcal{A}, \mathbf{r}) := \max_{\mathbf{s}, \mathbf{v} \in \mathcal{A}} \langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s} - \mathbf{v} \rangle$. The *width* of \mathcal{A} is the minimum directional width over all possible directions in its affine hull.

Pyramidal Directional Width. We define the pyramidal directional width of a set \mathcal{A} with respect to a direction \mathbf{r} and a base point $\mathbf{x} \in \mathcal{M}$ to be

$$PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x}) := \min_{\mathcal{S} \in \mathcal{S}_x} dirW(\mathcal{S} \cup \{\mathbf{s}(\mathcal{A}, \mathbf{r})\}, \mathbf{r}) = \min_{\mathcal{S} \in \mathcal{S}_x} \max_{\mathbf{s} \in \mathcal{A}, \mathbf{v} \in \mathcal{S}} \langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s} - \mathbf{v} \rangle, \quad (8)$$

where $\mathcal{S}_x := \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{A} \text{ such that } \mathbf{x} \text{ is a proper}^5 \text{ convex combination of all the elements in } \mathcal{S}\}$, and $\mathbf{s}(\mathcal{A}, \mathbf{r}) := \arg \max_{\mathbf{v} \in \mathcal{A}} \langle \mathbf{r}, \mathbf{v} \rangle$ is the FW atom used as a summit.

Pyramidal Width. To define the pyramidal width of a set, we take the minimum over the cone of possible *feasible* directions \mathbf{r} (in order to avoid the problem of zero width).

A direction \mathbf{r} is *feasible* for \mathcal{A} from \mathbf{x} if it points inwards $\text{conv}(\mathcal{A})$, (i.e. $\mathbf{r} \in \text{cone}(\mathcal{A} - \mathbf{x})$).

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

$$PWidth(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \text{faces}(\text{conv}(\mathcal{A})) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{r} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{0\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \mathbf{r}, \mathbf{x}). \quad (9)$$

Theorem 3. Let $\mathbf{x} \in \mathcal{M} = \text{conv}(\mathcal{A})$ be a suboptimal point and \mathcal{S} be an active set for \mathbf{x} . Let \mathbf{x}^* be an optimal point and corresponding error direction $\hat{\mathbf{e}} = (\mathbf{x}^* - \mathbf{x}) / \|\mathbf{x}^* - \mathbf{x}\|$, and negative gradient $\mathbf{r} := -\nabla f(\mathbf{x})$ (and so $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle > 0$). Let $\mathbf{d} = \mathbf{s} - \mathbf{v}$ be the pairwise FW direction obtained over \mathcal{A} and \mathcal{S} with negative gradient \mathbf{r} . Then

$$\frac{\langle \mathbf{r}, \mathbf{d} \rangle}{\langle \mathbf{r}, \hat{\mathbf{e}} \rangle} \geq PWidth(\mathcal{A}). \quad (10)$$

3.1 Properties of Pyramidal Width and Consequences

Examples of Values. The pyramidal width of a set \mathcal{A} is lower bounded by the minimal width over all subsets of atoms, and thus is strictly greater than zero if the number of atoms is finite. On the other hand, this lower bound is often too loose to be useful, as in particular, vertex subsets of the unit cube in dimension d can have exponentially small width $O(d^{-\frac{d}{2}})$ [see Corollary 27 in 36]. On the other hand, as we show here, the pyramidal width of the unit cube is actually $1/\sqrt{d}$, justifying why we kept the tighter but more involved definition (9). See Appendix B.1 for the proof.

Lemma 4. The pyramidal width of the unit cube in \mathbb{R}^d is $1/\sqrt{d}$.

For the probability simplex with d vertices, the pyramidal width is actually the same as its width, which is $2/\sqrt{d}$ when d is even, and $2/\sqrt{d-1}/d$ when d is odd [2] (see Appendix B.1). In contrast, the pyramidal width of an infinite set can be zero. For example, for a curved domain, the set of active atoms \mathcal{S} can contain vertices forming a very narrow pyramid, yielding a zero width in the limit.

Condition Number of a Set. The inverse of the rate constant ρ appearing in Theorem 1 is the product of two terms: L/μ is the standard *condition number* of the objective function appearing in the rates of gradient methods in convex optimization. The second quantity $(M/\delta)^2$ (diameter over pyramidal width) can be interpreted as a *condition number* of the domain \mathcal{M} , or its *eccentricity*. The more eccentric the constraint set (large diameter compared to its pyramidal width), the slower the convergence. The best condition number of a function is when its level sets are spherical; the analog in term of the constraint sets is actually the regular simplex, which has the maximum width-to-diameter ratio amongst all simplices [see Corollary 1 in 2]. Its eccentricity is (at most) $d/2$. In contrast, the eccentricity of the unit cube is d^2 , which is much worse.

⁵By *proper* convex combination, we mean that all coefficients are non-zero in the convex combination.

In the literature, the value L/σ is called the *condition number*.

It is easy to see that, the higher is the condition number, the slower be the convergence rate of our algorithm.

Define. A differentiable function $f(\mathbf{x})$ is called **smooth** iff it has a **Lipschitz continuous gradient**, i.e., iff $\exists L < \infty$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{z})\|_2 \leq L \|\mathbf{x} - \mathbf{z}\|_2, \quad \forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^N.$$

Lipschitz continuity of ∇f is a stronger condition than mere continuity, so any differentiable function whose gradient is Lipschitz continuous is in fact a **continuously differentiable** function.

We conjecture that the pyramidal width of a set of *vertices* (i.e. extrema of their convex hull) is *non-increasing* when another vertex is added (assuming that all previous points remain vertices). For example, the unit cube can be obtained by iteratively adding vertices to the regular probability simplex, and the pyramidal width thereby decreases from $2/\sqrt{d}$ to $1/\sqrt{d}$. This property could provide lower bounds for the pyramidal width of more complicated polytopes, such as $1/\sqrt{d}$ for the d -dimensional marginal polytope, as it can be obtained by removing vertices from the unit cube.

Complexity Lower Bounds. Combining the convergence Theorem 1 and the condition number of the unit simplex, we get a complexity of $O(d \frac{L}{\mu} \log(\frac{1}{\epsilon}))$ to reach ϵ -accuracy when optimizing a strongly convex function over the unit simplex. Here the linear dependence on d should not come as a surprise, in view of the known lower bound of $1/t$ for $t \leq d$ for Frank-Wolfe type methods [15].

Applications to Submodular Minimization. See Appendix A.2 for a consequence of our linear rate for the popular MNP algorithm for submodular function optimization (over the base polytope).

4 Non-Strongly Convex Generalization

Building on the work of Beck and Shtern [4] and Wang and Lin [33], we can generalize our global linear convergence results for all Frank-Wolfe variants for the more general case where $f(x) := g(Ax) + \langle b, x \rangle$, for $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^d$ and where g is μ_g -strongly convex and continuously differentiable over \mathcal{AM} . We note that for a general matrix A , f is convex but not necessarily *strongly* convex. In this case, the linear convergence still holds but with the constant μ appearing in the rate of Theorem 1 replaced with the generalized constant $\tilde{\mu}$ appearing in Lemma 9 in Appendix F.

5 Illustrative Experiments

We illustrate the performance of the presented algorithm variants in two numerical experiments, shown in Figure 2. The first example is a constrained Lasso problem (ℓ_1 -regularized least squares regression), that is $\min_{x \in \mathcal{M}} f(x) = \|Ax - b\|^2$, with $\mathcal{M} = 20 \cdot L_1$ a scaled L_1 -ball. We used a random Gaussian matrix $A \in \mathbb{R}^{200 \times 500}$, and a noisy measurement $b = Ax^*$ with x^* being a sparse vector with 50 entries ± 1 , and 10% of additive noise. For the L_1 -ball, the linear minimization oracle LMO_A just selects the column of A of best inner product with the residual vector. The second application comes from video co-localization. The approach used by [16] is formulated as a quadratic program (QP) over a flow polytope, the convex hull of paths in a network. In this application, the linear minimization oracle is equivalent to finding a shortest path in the network, which can be done easily by dynamic programming. For the LMO_A , we re-use the code provided by [16] and their included aeroplane dataset resulting in a QP over 660 variables. In both experiments, we see that the modified FW variants (away-steps and pairwise) outperform the original FW algorithm, and exhibit a linear convergence. In addition, the constant in the convergence rate of Theorem 1 can also be empirically shown to be fairly tight for AFW and PFW by running them on an increasingly obtuse triangle (see Appendix E).

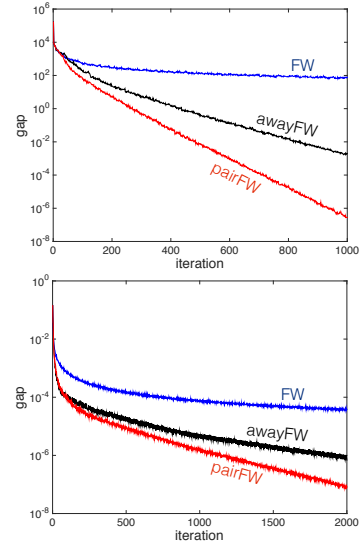
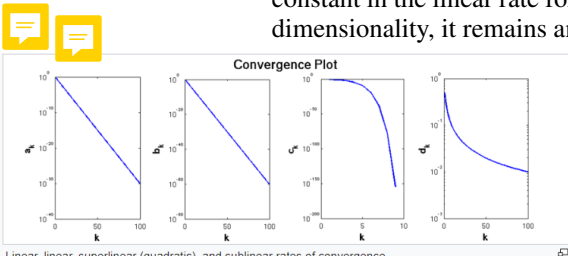


Figure 2: Duality gap g_t^{FW} vs iterations on the Lasso problem (top), and video co-localization (bottom). Code is available from the authors' website.

Discussion. Building on a preliminary version of our work [20], Beck and Shtern [4] also proved a linear rate for away-steps FW, but with a simpler lower bound for the LHS of (10) using linear duality arguments. However, their lower bound [see e.g. Lemma 3.1 in 4] is looser: they get a d^2 constant for the eccentricity of the regular simplex instead of the tighter d that we proved. Finally, the recently proposed generic scheme for *accelerating* first-order optimization methods in the sense of Nesterov from [24] applies directly to the FW variants given their global linear convergence rate that we proved. This gives for the first time first-order methods that *only use linear oracles* and obtain the “near-optimal” $\tilde{O}(1/k^2)$ rate for smooth convex functions, or the accelerated $\tilde{O}(\sqrt{L/\mu})$ constant in the linear rate for strongly convex functions. Given that the constants also depend on the dimensionality, it remains an open question whether this acceleration is practically useful.

Frank J.B. Alayrac, E. Hazan, A. Hubard, A. Osokin and P. Marcotte for helpful initially supported by the MSR-Inria Joint Center and a Google Research Award.

about convergence rates



Linear, linear, superlinear (quadratic), and sublinear rates of convergence

References

- [1] S. D. Ahipaaoglu, P. Sun, and M. Todd. Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.
- [2] R. Alexander. The width and diameter of a simplex. *Geometriae Dedicata*, 6(1):87–94, 1977.
- [3] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.
- [4] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *arXiv:1504.05002v1*, 2015.
- [5] A. Beck and M. 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.
- [6] M. D. Canon and C. D. Cullum. A tight upper bound on the rate of convergence of Frank-Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968.
- [7] V. Chari et al. On pairwise costs for network flow multi-object tracking. In *CVPR*, 2015.
- [8] J. 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.
- [9] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.
- [10] D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv:1301.4666v5*, 2013.
- [11] D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *ICML*, 2015.
- [12] J. Guélat and P. Marcotte. Some comments on Wolfe’s ‘away step’. *Mathematical Programming*, 1986.
- [13] D. Hearn, S. Lawphongpanich, and J. Ventura. Restricted simplicial decomposition: Computation and extensions. In *Computation Mathematical Programming*, volume 31, pages 99–118. Springer, 1987.
- [14] C. A. Holloway. An extension of the Frank and Wolfe method of feasible directions. *Mathematical Programming*, 6(1):14–27, 1974.
- [15] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.
- [16] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *ECCV*, 2014.
- [17] V. Kolmogorov and R. Zabini. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.
- [18] R. G. Krishnan, S. Lacoste-Julien, and D. Sontag. Barrier Frank-Wolfe for marginal inference. In *NIPS*, 2015.
- [19] P. Kumar and E. A. Yildirim. A linearly convergent linear-time first-order algorithm for support vector classification with a core set result. *INFORMS Journal on Computing*, 2010.
- [20] S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *arXiv:1312.7864v2*, 2013.
- [21] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013.
- [22] G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv:1309.5550v2*, 2013.
- [23] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):787–823, Jan. 1966.
- [24] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.
- [25] B. Mitchell, V. F. Demyanov, and V. Malozemov. Finding the point of a polyhedron closest to the origin. *SIAM Journal on Control*, 12(1), 1974.
- [26] R. Nanculef, E. Frandi, C. Sartori, and H. Allende. A novel Frank-Wolfe algorithm. Analysis and applications to large-scale SVM training. *Information Sciences*, 2014.
- [27] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.
- [28] J. Pena, D. Rodriguez, and N. Soheili. On the von Neumann and Frank-Wolfe algorithms with away steps. *arXiv:1507.04073v2*, 2015.
- [29] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, pages 185–208. 1999.
- [30] S. M. Robinson. *Generalized Equations and their Solutions, Part II: Applications to Nonlinear Programming*. Springer, 1982.
- [31] B. Von Hohenbalken. Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming*, 13(1):49–68, 1977.
- [32] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(12):1–305, 2008.
- [33] P.-W. Wang and C.-J. Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.
- [34] P. Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*. 1970.
- [35] P. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.
- [36] G. M. Ziegler. Lectures on 0/1-polytopes. *arXiv:math/9909177v1*, 1999.

Appendix

Outline. The appendix is organized as follows: In Appendix A, we discuss some of the Frank-Wolfe algorithm variants in more details and related work (including the MNP algorithm in Appendix A.1 and its application to submodular minimization in Appendix A.2).

In Appendix B, we discuss the pyramidal width for some particular cases of sets (such as the probability simplex and the unit cube), and then provide the proof of the main Theorem 3 relating the pyramidal width to the progress quantity essential for the linear convergence rate. Section C presents an affine invariant version of the complexity constants.

In the following Section D, we show the main linear convergence result for the four variants of the FW algorithm, and also discuss the sublinear rates for general convex functions. Section E presents a simple experiment demonstrating the empirical tightness of the theoretical linear convergence rate constant. Finally, in Appendix F we discuss the generalization of the linear convergence to some cases of non-strongly convex functions in more details.

A More on Frank-Wolfe Algorithm Variants

A.1 Wolfe’s Min-Norm Point (MNP) algorithm

A generalization of Wolfe’s min-norm point (MNP) algorithm [35] for general convex functions is to run Algorithm 3 with the correction subroutine in step 7 implemented as presented below in Algorithm 5. In Wolfe’s paper [35], the correction step is called the minor cycle; whereas the FW outer loop is called the major cycle.

As we have mentioned in Section 1, MNP for polytope distance is often confused with fully-corrective FW as presented in Algorithm 3, for quadratic objectives. In fact, standard FCFW optimizes f over $\text{conv}(\mathcal{A}^{(t)})$, whereas MNP implements the correction as a sequence of *affine* projections on the active set that potentially yield a different update.

Algorithm 5 Generalized version of Wolfe’s MNP correction: **MNP-Correction**($\mathbf{x}^{(t)}, \mathcal{A}^{(t)}, \mathbf{s}_t$)

```

1: Let  $\mathcal{S}^{(0)} := \mathcal{A}^{(t)} \cup \{\mathbf{s}_t\}$ , and  $\mathbf{z}_0 := \mathbf{x}^{(t)}$ . Note that  $\mathbf{x}^{(t)} = \sum_{\mathbf{v} \in \mathcal{A}^{(t)}} \alpha_{\mathbf{v}} \mathbf{v}$  and we assume that
   the elements of  $\mathcal{A}^{(t)}$  are affinely independent.
2: for  $k = 1 \dots |\mathcal{S}^{(0)}|$  do
3:   Let  $\mathbf{y}_k$  be the minimizer of  $f$  on the affine hull of  $\mathcal{S}^{(k-1)}$ 
4:   if  $\mathbf{y}_k$  is in the relative interior of  $\text{conv}(\mathcal{S}^{(k-1)})$  then
5:     return  $(\mathbf{y}_k, \mathcal{S}^{(k-1)})$  ( $\mathcal{S}^{(k-1)}$  is active set for  $\mathbf{y}_k$ )
6:   else
7:     Let  $\mathbf{z}_k$  be the solution of doing line-search from  $\mathbf{z}_{k-1}$  to  $\mathbf{y}_k$ . [step 2(d)(iii) of Alg. 1 in 39]
8:     (Note that  $\mathbf{z}_k$  now lies on the boundary of  $\text{conv}(\mathcal{S}^{(k-1)})$ , and so some atoms were removed)
9:     Let  $\mathcal{S}^{(k)}$  be the (affinely independent) active atoms in the expansion of  $\mathbf{z}_k$ .
10:   end if
11: end for

```

There are two main differences between FCFW and the MNP algorithm. First, after a correction step, MNP guarantees that $\mathbf{x}^{(t+1)}$ is *both* the minimizer of f over the *affine hull* of $\mathcal{A}^{(t+1)}$ and also $\text{conv}(\mathcal{A}^{(t+1)})$ (where $\mathcal{A}^{(t+1)}$ might be much smaller than $\mathcal{A}^{(t)} \cup \{\mathbf{s}_t\}$), whereas FCFW guarantees that $\mathbf{x}^{(t+1)}$ is the minimizer of f over $\text{conv}(\mathcal{A}^{(t)} \cup \{\mathbf{s}_t\})$ – this is usually not the case for MNP unless at most one atom was dropped from the correction polytope, as is apparent from our convergence proof. Secondly, the correction atoms $\mathcal{A}^{(t)}$ are always affinely independent for MNP and are identical to the active set $\mathcal{S}^{(t)}$, whereas FCFW can use both redundant as well as inactive atoms. The advantage of the MNP implementation using affine hull projections is that the correction can be efficiently implemented when f is the Euclidean norm, especially when a triangular array repre-

sensation of the active set is maintained (see the careful implementation details in Wolfe’s original paper [35]).

The MNP variant indeed only makes sense when the minimization of f over the affine hull of \mathcal{M} is well-defined (and is efficient). Note though that the line-search in step 7 does not require any new information about \mathcal{M} , as it is made only with respect to $\text{conv}(\mathcal{S}^{(k-1)})$, for which we have an explicit list of vertices. This line-search can be efficiently computed in $O(|\mathcal{S}^{(k-1)}|)$, and is well described for example in step 2(d)(iii) of Algorithm 1 of Chakrabarty et al. [39].

A.2 Applications to Submodular Minimization

An interesting consequence of our global linear convergence result for FW algorithm variants here is the potential to reduce the gap between the known theoretical rates and the impressive empirical performance of MNP for submodular function minimization (over the base polytope). While Bach [3] already showed convergence of FW in this case, Chakrabarty et al. [39] later gave a weaker convergence rate for Wolfe’s MNP variant. For exact submodular function optimization, the overall complexity by [39] was $O(d^5 F^2)$ (with some corrections⁶), where F is the maximum absolute value of the integer-valued submodular function. This is in contrast to $O(d^5 \log(d F))$ for the fastest algorithms [40]. Using our linear convergence, the F factor can be put back in the log term for MNP,⁷ matching their empirical observations that the MNP algorithm was not too sensitive to F . The same follows for AFW and FCFW, which is novel.

A.3 Pairwise Frank-Wolfe

Our new analysis of the pairwise Frank-Wolfe variant as introduced in Section 1 is motivated by the work of Garber and Hazan [10], who provided the first variant of Frank-Wolfe with a global linear convergence rate with explicit constants that do not depend on the location of the optimum \mathbf{x}^* , for a more complex extension of such a pairwise algorithm. An important contribution of the work of Garber and Hazan [10] was to define the concept of *local linear oracle*, which (approximately) minimizes a linear function on the intersection of \mathcal{M} and a small ball around $\mathbf{x}^{(t)}$ (hence the name *local*). They showed that if such a local linear oracle was available, then one could replace the step that moves towards \mathbf{s}_t in the standard FW procedure with a constant step-size move towards the point returned by the local linear oracle to obtain a globally linearly convergent algorithm. They then demonstrated how to implement such a local linear oracle by using only one call to the linear oracle (to get \mathbf{s}_t), as well as sorting the atoms in $\mathcal{S}^{(t)}$ in decreasing order of their inner product with $\nabla f(\mathbf{x}^{(t)})$ (note that the first element then is the away atom \mathbf{v}_t from Algorithm 1). The procedure implementing the local linear oracle amounts to iteratively swapping the mass from the away atom \mathbf{v}_t to the FW atom \mathbf{s}_t until enough mass has been moved (given by some precomputed constants). If the amount of mass to move is bigger than $\alpha_{\mathbf{v}_t}^{(t)}$, then one sets $\alpha_{\mathbf{v}_t}^{(t)}$ to zero and start moving mass from the *second* away atom, and so on, until enough mass has been moved (which is why the sorting is needed). We call such a swap of mass between the away atom and the FW atom a *pairwise FW* step, i.e. $\alpha_{\mathbf{v}_t}^{(t+1)} = \alpha_{\mathbf{v}_t}^{(t)} - \gamma$ and $\alpha_{\mathbf{s}_t}^{(t+1)} = \alpha_{\mathbf{s}_t}^{(t)} + \gamma$ for some step-size $\gamma \leq \gamma_{\max} := \alpha_{\mathbf{v}_t}^{(t)}$. The local linear oracle is implemented as a sequence of pairwise FW steps, always keeping the same FW atom \mathbf{s}_t as the target, but updating the away atom to move from as we set their coordinates to zero.

A major disadvantage of the algorithm presented by Garber and Hazan [10] is that their algorithm is *not adaptive*: it requires the computation of several (loose) constants to determine the step-sizes, which means that the behavior of the algorithm is stuck in its worst-case analysis. The pairwise Frank-Wolfe variant is obtained by simply doing one line-search in the pairwise Frank-Wolfe direction $\mathbf{d}_t^{\text{PFW}} := \mathbf{s}_t - \mathbf{v}_t$ (see Algorithm 2). This gives a fully adaptive algorithm, and it turns out that this is sufficient to yield a global linear convergent rate.

⁶Chakrabarty et al. [39] quoted a complexity of $O(d^7 F^2)$ for MNP. However, this fell short of the earlier result of Bach [3] for classic FW in the submodular minimization case, which was better by two $O(d)$ factors. [39] counted $O(d^3)$ per iteration of the MNP algorithm whereas Wolfe had provided a $O(d^2)$ implementation; and they missed that there were at least $t/2$ good cycles (‘non drop steps’) after t iterations, rather than $O(t/d)$ as they have used.

⁷This is assuming that the eccentricity of the base polytope does not depend on F , which remains to be proven.

Notes on Convergence Proofs in [26]. We here point out some corrections to the convergence proofs given in [26] for a variant of pairwise FW that chooses between a standard FW step and a pairwise FW step by picking the one which makes the most progress on the objective after a line-search. [26, Proposition 1] states the global convergence of their algorithm by arguing that $\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t^{\text{PFW}} \rangle \geq \langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t^{\text{FW}} \rangle$ and then stating that they can re-use the same pattern as the standard FW convergence proof but with the direction $\mathbf{d}_t^{\text{PFW}}$. But this is forgetting the fact that the maximal step-size $\gamma_{\max} = \alpha_{v_t}$ for a pairwise FW step can be too small to make sufficient progress. Their global convergence statement is still correct as every step of their algorithm makes more progress than a FW step, which already has a global convergence result, but this is not the argument they made. Similarly, they state a global linear convergence result in their Proposition 4, citing a proof from [37]. On the other hand, the relevant used Proposition 3 in [37] forgot to consider the possibility of problematic *swap steps* that we had to painfully bound in our convergence Theorem 8; they only considered drop steps or ‘good steps’, thereby missing a bound on the number of swap steps to get a valid global bound.

A.4 Other Related Work

Very recently, following the earlier workshop version of our article [20], Pena et al. [28] presented an alternative geometric quantity measuring the linear convergence speed of the AFW algorithm variant. Their approach is motivated by a special case of the Frank-Wolfe method, the von Neumann algorithm. Their complexity constant – called the restricted width – is also bounded away from zero, but its value does depend on the location of the optimal solution, which is a disadvantage shared with the earlier existing results of [34, 12, 5], as well as the line of work of [1, 19, 26] that relies on Robinson’s condition [30]. More precisely, the bound on the constant given in [28, Theorem 4] applies to the translated atoms \tilde{A} relative to the optimum point. The constant is not affine-invariant, whereas the constants μ_f^A (22) and C_f^A (26) in our setting are so, see the discussion in Section C. It would still be interesting to compare the value of our respective constants on standard polytopes.

B Pyramidal Width

B.1 Pyramidal Width of the Cube and Probability Simplex

Lemma’ 4. *The pyramidal width of the unit cube in \mathbb{R}^d is $1/\sqrt{d}$.*

Proof of Lemma 4. First consider a point \mathbf{x} in the interior of the cube, and let \mathbf{r} be the unit length direction achieving the smallest pyramidal width for \mathbf{x} . Let $\mathbf{s} = \mathbf{s}(\mathcal{A}, \mathbf{r})$ (the FW atom in direction \mathbf{r}). Without loss of generality, by symmetry,⁸ we can rotate the cube so that \mathbf{s} lies at the origin. This implies that each coordinate of \mathbf{r} is non-positive. Represent a vertex \mathbf{v} of the cube as its set of indices for which $v_i = 1$. Then $\langle \mathbf{r}, \mathbf{s} - \mathbf{v} \rangle = \sum_{i \in \mathbf{v}} -r_i \geq \max_{i \in \mathbf{v}} |r_i|$. Consider any possible active set \mathcal{S} ; as \mathbf{x} has all its coordinate strictly positive, for each dimension i , there must exist an element of \mathcal{S} with its i coordinate equals to 1. This means that $\max_{\mathbf{v} \in \mathcal{S}} \langle \mathbf{r}, \mathbf{s} - \mathbf{v} \rangle \geq \|\mathbf{r}\|_\infty$. But as \mathbf{r} has unit Euclidean norm, then $\|\mathbf{r}\|_\infty \geq 1/\sqrt{d}$. Now consider \mathbf{x} to lie on a facet of the cube (i.e. the active set \mathcal{S} is lower dimensional); and let $I := \{i : r_i < 0\}$. Since \mathbf{r} has to be feasible from \mathbf{x} , for each $i \in I$, we cannot have $x_i = 0$ and thus there exists an element of the active set with its i^{th} coordinate equal to 1. We thus have that $\max_{\mathbf{v} \in \mathcal{S}} \langle \mathbf{r}, \mathbf{s} - \mathbf{v} \rangle \geq \|\mathbf{r}\|_\infty \geq 1/\sqrt{|I|} \geq 1/\sqrt{d}$. Using the same argument on a lower dimensional \mathcal{K} give a lower bound of $1/\sqrt{\dim(\mathcal{K})}$ which is bigger. These cover all the possibilities appearing in the definition of the pyramidal width, and thus the lower bound is correct. It is achieved by choosing an \mathbf{x} in the interior, the canonical basis as the active set \mathcal{S} , and the direction defined by $r_i = -1/\sqrt{d}$ for each i . \square

We note that both the active set definition \mathcal{S} and the feasibility condition on \mathbf{r} were crucially used in the above proof to obtain such a large value for the pyramidal width of the unit cube, thus justifying the somewhat involved definition appearing in (9). On the other hand, the astute reader might have noticed that the important quantity to lower bound for the linear convergence rate of the different FW variants is $\frac{\langle \mathbf{r}_t, \tilde{\mathbf{d}}_t \rangle}{\langle \mathbf{r}_t, \tilde{\mathbf{e}}_t \rangle}$ (as in (5)), rather than the looser value $\frac{1}{M} \frac{\langle \mathbf{r}_t, \mathbf{d}_t \rangle}{\langle \mathbf{r}_t, \mathbf{e}_t \rangle}$ that we used to handle the

⁸We thank Jean-Baptiste Alayrac for inspiring us to use symmetry in the proof.

proof of the difficult Theorem 3 (where we recall that M is the diameter of \mathcal{M}). One could thus hope to get a tighter measure for the condition number of a set by considering $\|s - v\|$ (with s and v the minimizing witnesses for the pyramidal width) instead of the diameter M in the ratio diameter / pyramidal width. This might give a tighter constant for general sets, but in the case of the cube, it does not change the general $\Omega(d^2)$ dependence for its condition number. To see this, suppose that d is even and let $k = d/2$. Consider the direction r with $r_i := -1$ for $1 \leq i \leq k$, and $r_i := -\epsilon$ for $(k+1) \leq i \leq d$. We thus have that the FW atom $s(\mathcal{A}, r)$ is the origin as before. Consider x such that $x_i := 1/k$ for $1 \leq i \leq k$, and $x_i := 1$ for $(k+1) \leq i \leq d$, that is, x is the uniform convex combination of the k vertices which has only one non-zero in the first k coordinates, and the last k coordinates all equal to 1. We have that r is a feasible direction from x , and that all vertices in the active set for x have the same inner product with r : $\max_{v \in \mathcal{S}} \langle r, s - v \rangle = 1 + k\epsilon$. We thus have:

$$\left\langle \frac{r}{\|r\|}, \frac{s - v}{\|s - v\|} \right\rangle = \frac{1 + k\epsilon}{(\sqrt{k}\sqrt{1 + \epsilon^2})(\sqrt{k+1})} \leq \frac{1}{k} \quad \text{for } \epsilon \text{ small enough.}$$

Squaring the inverse, we thus get that the condition number of the cube is at least $k^2 = d^2/4$ even using this tighter definition, thus not changing the $\Omega(d^2)$ dependence.

Pyramidal Width for the Probability Simplex. For any x in the relative interior of the probability simplex on d vertices, we have that $\mathcal{S} = \mathcal{A}$, and thus the pyramidal directional width (8) in the feasible direction r with base point x is the same as the standard directional width. Moreover, any face of the probability simplex is just a probability simplex in lower dimensions (with bigger width). This is why the pyramidal width of the probability simplex is the same as its standard width. The width of a regular simplex was implicitly given in [2]; we provide more details here on this citation. Alexander [2] considers a regular simplex with k vertices and side length Δ . For any partition of the k points into a set of r and $k - r$ points (for $r \leq \lfloor k/2 \rfloor$), one can compute the distance $c(r, \Delta)$ between the flats (affine hulls) of the two sets (which also corresponds to a specific directional width). Alexander gives a formula on the third line of p. 91 in [2] for the square of this distance:

$$(c(r, \Delta))^2 = \Delta^2 \frac{k}{2r(k - r)}. \quad (11)$$

The width of the regular simplex is obtained by taking the minimum of (11) with respect to $r \leq \lfloor k/2 \rfloor$. As (11) is a decreasing function up to $r = k/2$, we obtain its minimum by substituting $r = \lfloor k/2 \rfloor$. By using $\Delta = \sqrt{2}$, $k = d$ and $r = \lfloor d/2 \rfloor$ in (11), we get that the width for the probability simplex on d vertices is $2/\sqrt{d}$ when d is even and the slightly bigger $2/\sqrt{d - 1/d}$ when d is odd.

From the pyramidal width perspective, one can obtain these numbers by considering any relative interior point of the probability simplex as x and considering the following feasible r . For d even, we let $r_i := 1$ for $1 \leq i \leq d/2$ and $r_i := -1$ for $i > d/2$. Note that $\sum_i r_i = 0$ and thus r is a feasible direction for a point in the relative interior of the probability simplex. Then $\max_{s, v \in \mathcal{A}} \langle \frac{r}{\|r\|}, s - v \rangle = \frac{1}{\sqrt{d}}(1 + 1) = \frac{2}{\sqrt{d}}$ as claimed. For d odd, we can choose $r_i := \frac{2}{d-1}$ for $1 \leq i \leq \frac{d-1}{2}$, and $r_i := \frac{2}{d+1}$ for $i \geq \frac{d+1}{2}$. Then $\|r\| = \sqrt{\frac{4d}{d^2-1}}$ and $\max_{s, v \in \mathcal{A}} \langle \frac{r}{\|r\|}, s - v \rangle = \sqrt{\frac{4d}{d^2-1}} = 2/\sqrt{d - 1/d}$ as claimed. Showing that these obtained values were the minimum possible ones is non-trivial though, which is why we appealed to the width of the regular simplex computed from [2].

B.2 Proof of Theorem 3 on the Pyramidal Width

In this section, we prove the main technical result in our paper: a geometric lower bound for the crucial quantity appearing in the linear convergence rate for the FW optimization variants.

Theorem' 3. Let $x \in \mathcal{M} = \text{conv}(\mathcal{A})$ be a suboptimal point and \mathcal{S} be an active set for x . Let x^* be an optimal point and corresponding error direction $\hat{e} = (x^* - x)/\|x^* - x\|$, and negative gradient $r := -\nabla f(x)$ (and so $\langle r, \hat{e} \rangle > 0$). Let d^{PFW} be the pairwise FW direction obtained over \mathcal{A} and \mathcal{S} with negative gradient r . Then we have:

$$\frac{\langle r, d^{\text{PFW}} \rangle}{\langle r, \hat{e} \rangle} \geq \text{PWidth}(\mathcal{A}). \quad (12)$$

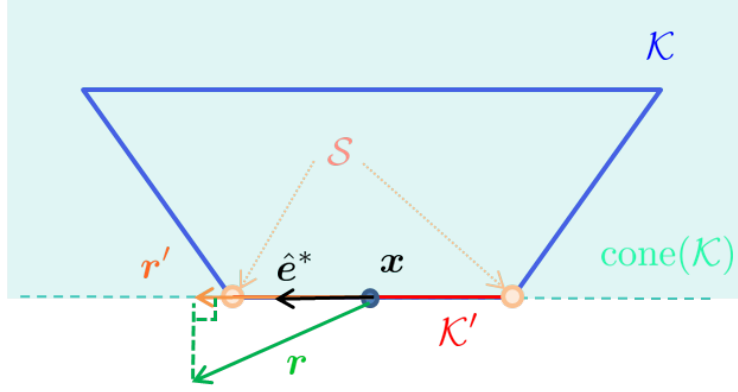


Figure 3: Depiction of the quantities in the proof of Lemma 5. If $\mathbf{r} \notin \text{cone}(\mathcal{K})$, but $\mathbf{r} \in \text{span}(\mathcal{K})$, then the unit vector direction $\hat{\mathbf{e}}^*$ minimizing the angle with \mathbf{r} is generated by a point \mathbf{x}^* lying on a facet \mathcal{K}' of the polytope \mathcal{K} that contains \mathbf{x} .

We first give a proof sketch, and then give the full proof.

Recall that a direction \mathbf{r} is *feasible* for \mathcal{A} from \mathbf{x} if it points inwards $\text{conv}(\mathcal{A})$, i.e. $\mathbf{r} \in \text{cone}(\mathcal{A} - \mathbf{x})$.

Warm-Up Proof Sketch. By Cauchy-Schwarz, the denominator of (12) is at most $\|\mathbf{r}\|$. If \mathbf{r} is a feasible direction from \mathbf{x} for \mathcal{A} , then the LHS is lower bounded by $PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x})$ as \mathbf{d}^{PFW} is included as a possible $\mathbf{s} - \mathbf{v}$ direction considered in the definition of $PdirW$ (8). If \mathbf{r} is not feasible from \mathbf{x} , this means that \mathbf{x} lies on the boundary of \mathcal{M} . One can then show that the potential \mathbf{x}^* that can maximize $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle$ has also to lie on a facet \mathcal{K} of \mathcal{M} containing \mathbf{x} (see Lemma 5 below). The idea is then to project \mathbf{r} onto \mathcal{K} , and re-do the argument with \mathcal{K} replacing \mathcal{M} and show that the inequality is in the right direction. This explains why all the subfaces of \mathcal{M} are considered in the definition of the pyramidal width (9), and that only *feasible* directions are considered. \square

Lemma 5 (Minimizing angle is on a facet). *Let \mathbf{x} be at the origin, inside a polytope \mathcal{K} and suppose that $\mathbf{r} \in \text{span}(\mathcal{K})$ is not a feasible direction for \mathcal{K} from \mathbf{x} (i.e. $\mathbf{r} \notin \text{cone}(\mathcal{K})$). Then a feasible direction in \mathcal{K} minimizing the angle with \mathbf{r} lies on a facet⁹ \mathcal{K}' of \mathcal{K} that includes the origin \mathbf{x} . That is:*

$$\max_{\mathbf{e} \in \mathcal{K}} \left\langle \mathbf{r}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \right\rangle = \max_{\mathbf{e} \in \mathcal{K}'} \left\langle \mathbf{r}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \right\rangle = \max_{\mathbf{e} \in \mathcal{K}'} \left\langle \mathbf{r}', \frac{\mathbf{e}}{\|\mathbf{e}\|} \right\rangle \quad (13)$$

where \mathcal{K}' contains \mathbf{x} , $\|\cdot\|$ is the Euclidean norm and \mathbf{r}' is defined as the orthogonal projection of \mathbf{r} on $\text{span}(\mathcal{K}')$.

Proof. This seems like an obvious geometric fact (see Figure 3), but we prove it formally, as sometimes high dimensional geometry is tricky (for example, the result is false without the assumption that $\mathbf{r} \in \text{span}(\mathcal{K})$ or if $\|\cdot\|$ is not the Euclidean norm). Rewrite the optimization variable on the LHS of (13) as $\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|}$. The optimization domain for $\hat{\mathbf{e}}$ is thus the intersection between the unit sphere and $\text{cone}(\mathcal{K})$. We now show that any maximizer $\hat{\mathbf{e}}^*$ cannot lie in the relative interior of $\text{cone}(\mathcal{K})$, and thus it has to lie on a *facet* of $\text{cone}(\mathcal{K})$, implying then that a corresponding maximizer \mathbf{e}^* is lying on a facet of \mathcal{K} containing \mathbf{x} , concluding the proof for the first equality in (13).

First, as $\mathbf{r} \in \text{span}(\mathcal{K})$, we can consider without loss of generality that $\text{cone}(\mathcal{K})$ is full dimensional by projecting on its affine hull if needed. We want to solve $\max_{\hat{\mathbf{e}}} \langle \mathbf{r}, \hat{\mathbf{e}} \rangle$ s.t. $\|\hat{\mathbf{e}}\|^2 = 1$ and $\hat{\mathbf{e}} \in \text{cone}(\mathcal{K})$.

⁹As a reminder, we define a *k-face* of \mathcal{M} (a *k-dimensional face* of \mathcal{M}) a set \mathcal{K} such that $\mathcal{K} = \mathcal{M} \cap \{\mathbf{y} : \langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle = 0\}$ for some normal vector \mathbf{r} and fixed reference point $\mathbf{x} \in \mathcal{K}$ with the additional property that \mathcal{M} lies on one side of the given half-space determined by \mathbf{r} i.e. $\langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle \leq 0 \forall \mathbf{y} \in \mathcal{M}$. k is the dimensionality of the affine hull of \mathcal{K} . We call a *k-face* of dimensions $k = 0, 1, \dim(\mathcal{M}) - 2$ and $\dim(\mathcal{M}) - 1$ a *vertex*, *edge*, *ridge* and *facet* respectively. \mathcal{M} is a *k-face* of itself with $k = \dim(\mathcal{M})$. See Definition 2.1 in the book of Ziegler [41], which we also recommend for more background material on polytopes.

By contradiction, we suppose that \hat{e}^* lies in the interior of $\text{cone}(\mathcal{K})$, and so we can remove the polyhedral cone constraint. The gradient of the objective is the constant \mathbf{r} and the gradient of the equality constraint is $2\hat{e}$. By the Karush-Kuhn-Tucker (KKT) necessary conditions for a stationary point to the problem with the only equality constraint $\|\hat{e}\|^2 = 1$ (see e.g. [Proposition 3.3.1 in 38]), then the gradient of the objective is collinear to the gradient of the equality constraint, i.e. we have $\hat{e}^* = \pm \hat{\mathbf{r}}$. Since $\hat{\mathbf{r}}$ is not feasible, then $\hat{e}^* = -\hat{\mathbf{r}}$, which is actually a local *minimum* of the inner product by Cauchy-Schwarz. We thus conclude that the maximizing \hat{e}^* lies on the boundary of $\text{cone}(\mathcal{K})$, concluding the proof for the first equality in (13).

For the second equality in (13), we simply use the fact that $\mathbf{r} - \mathbf{r}'$ is orthogonal to the elements of \mathcal{K}' by the definition of the orthogonal projection. \square

Proof of Theorem 3. Let $\hat{e}(\mathbf{x}^*) := \frac{\mathbf{x}^* - \mathbf{x}}{\|\mathbf{x}^* - \mathbf{x}\|}$ be the normalized error vector. We consider the worst-case possibility for \mathbf{x}^* . As \mathbf{x} is not optimal, we require that $\langle \mathbf{r}, \hat{e}(\mathbf{x}^*) \rangle > 0$. We recall that by definition of the pairwise FW direction:

$$\left\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{d}^{\text{PFW}} \right\rangle = \max_{s \in \mathcal{A}, v \in \mathcal{S}} \left\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, s - v \right\rangle \geq \min_{S' \in \mathcal{S}_x} \max_{s \in \mathcal{A}, v \in S'} \left\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, s - v \right\rangle = \text{PdirW}(\mathcal{A}, \mathbf{r}, \mathbf{x}). \quad (14)$$

By Cauchy-Schwarz, we always have $\langle \mathbf{r}, \hat{e}(\mathbf{x}^*) \rangle \leq \|\mathbf{r}\|$. If \mathbf{r} is a feasible direction from \mathbf{x} in $\text{conv}(\mathcal{A})$, then \mathbf{r} appears in the set of directions considered in the definition of the pyramidal width (9) for \mathcal{A} and so from (14), we have that the inequality (12) holds.

If \mathbf{r} is not a feasible direction, then we iteratively project it on the faces of \mathcal{M} until we get a feasible direction \mathbf{r}' , obtaining a term $\text{PdirW}(\mathcal{A} \cap \mathcal{K}, \mathbf{r}', \mathbf{x})$ for some face \mathcal{K} of \mathcal{M} as appearing in the definition of the pyramidal width (9). The rest of the proof formalizes this process. As \mathbf{x} is fixed, we work on the centered polytope at \mathbf{x} to simplify the statements, i.e. let $\tilde{\mathcal{M}} := \mathcal{M} - \mathbf{x}$. We have the following worst case lower bound for (12):

$$\frac{\langle \mathbf{r}, \mathbf{d}^{\text{PFW}} \rangle}{\langle \mathbf{r}, \hat{e} \rangle} \geq \left(\max_{s \in \mathcal{A}, v \in \mathcal{S}} \langle \mathbf{r}, s - v \rangle \right) \left(\max_{e \in \tilde{\mathcal{M}}} \langle \mathbf{r}, \frac{e}{\|e\|} \rangle \right)^{-1}. \quad (15)$$

The first term on the RHS of (15) just comes from the definition of \mathbf{d}^{PFW} (with equality), whereas the second term is considering the worst case possibility for \mathbf{x}^* to lower bound the LHS. Note also that the second term has to be strictly greater to zero since \mathbf{x} is not optimal.

Without loss of generality, we can assume that $\mathbf{r} \in \text{span}(\tilde{\mathcal{M}})$ (otherwise, just project it), as any orthogonal component would not change the inner products appearing in (15). If (this projected) \mathbf{r} is feasible from \mathbf{x} , then $\max_{e \in \tilde{\mathcal{M}}} \langle \mathbf{r}, \frac{e}{\|e\|} \rangle = \|\mathbf{r}\|$, and we again have the lower bound (14) arising in the definition of the pyramidal width.

We thus now suppose that \mathbf{r} is not feasible. By the Lemma 5, we have the existence of a facet \mathcal{K}' of $\tilde{\mathcal{M}}$ that includes the origin \mathbf{x} such that:

$$\max_{e \in \tilde{\mathcal{M}}} \left\langle \mathbf{r}, \frac{e}{\|e\|} \right\rangle = \max_{e \in \mathcal{K}'} \left\langle \mathbf{r}, \frac{e}{\|e\|} \right\rangle = \max_{e \in \mathcal{K}'} \left\langle \mathbf{r}', \frac{e}{\|e\|} \right\rangle, \quad (16)$$

where \mathbf{r}' is the result of the orthogonal projection of \mathbf{r} on $\text{span}(\mathcal{K}')$. We now look at how the numerator of (15) transforms when considering \mathbf{r}' and \mathcal{K}' :

$$\begin{aligned} \max_{s \in \mathcal{A}, v \in \mathcal{S}} \langle \mathbf{r}, s - v \rangle &= \max_{s \in \mathcal{M}} \langle \mathbf{r}, s - \mathbf{x} \rangle + \max_{v \in \mathcal{S}} \langle -\mathbf{r}, v - \mathbf{x} \rangle \\ &\geq \max_{s \in (\mathcal{K}' + \mathbf{x})} \langle \mathbf{r}, s - \mathbf{x} \rangle + \max_{v \in \mathcal{S} \cap (\mathcal{K}' + \mathbf{x})} \langle -\mathbf{r}, v - \mathbf{x} \rangle \\ &= \max_{s \in (\mathcal{K}' + \mathbf{x})} \langle \mathbf{r}', s - \mathbf{x} \rangle + \max_{v \in \mathcal{S}} \langle -\mathbf{r}', v - \mathbf{x} \rangle \\ &= \max_{s \in \mathcal{A} \cap (\mathcal{K}' + \mathbf{x}), v \in \mathcal{S}} \langle \mathbf{r}', s - v \rangle. \end{aligned} \quad (17)$$

To go from the first to the second line, we use the fact that the first term yields an inequality as $(\mathcal{K}' + \mathbf{x}) \subseteq (\tilde{\mathcal{M}} + \mathbf{x}) = \mathcal{M}$. Also, since \mathbf{x} is in the relative interior of $\text{conv}(\mathcal{S})$ (as \mathbf{x} is a *proper* convex combination of elements of \mathcal{S} by definition), we have that $(\mathcal{S} - \mathbf{x}) \subseteq \mathcal{K}$ for any face \mathcal{K} of $\tilde{\mathcal{M}}$

containing the origin \mathbf{x} . Thus $\mathcal{S} = \mathcal{S} \cap (\mathcal{K}' + \mathbf{x})$, and the second term on the first line actually yields an equality for the second line. The third line uses the fact that $\mathbf{r} - \mathbf{r}'$ is orthogonal to members of \mathcal{K}' , as \mathbf{r}' is obtained by orthogonal projection.

Plugging (16) and (17) into the inequality (15), we get:

$$\frac{\langle \mathbf{r}, \mathbf{d}^{\text{PFW}} \rangle}{\langle \mathbf{r}, \hat{\mathbf{e}} \rangle} \geq \left(\max_{\substack{\mathbf{s} \in \mathcal{A} \cap (\mathcal{K}' + \mathbf{x}), \\ \mathbf{v} \in \mathcal{S}}} \langle \mathbf{r}', \mathbf{s} - \mathbf{v} \rangle \right) \left(\max_{\mathbf{e} \in \mathcal{K}'} \langle \mathbf{r}', \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle \right)^{-1}. \quad (18)$$

We are back to a similar situation to (15), with the lower dimensional \mathcal{K}' playing the role of the polytope $\tilde{\mathcal{M}}$, and $\mathbf{r}' \in \text{span}(\mathcal{K}')$ playing the role of \mathbf{r} . If \mathbf{r}' is feasible from \mathbf{x} in \mathcal{K}' , then re-using the previous argument, we get $\text{Pdir}W(\mathcal{A} \cap (\mathcal{K}' + \mathbf{x}), \mathbf{r}', \mathbf{x})$ as the lower bound, which is part of the definition of the pyramidal width of \mathcal{A} (note that we have $(\mathcal{K}' + \mathbf{x})$ as \mathcal{K}' is a face of the *centered* polytope $\tilde{\mathcal{M}}$). Otherwise (if $\mathbf{r} \notin \text{cone}(\mathcal{K}')$), then we use Lemma 5 again to get a facet \mathcal{K}'' of \mathcal{K}' as well as a new direction \mathbf{r}'' which is the orthogonal projection of \mathbf{r}' on $\text{span}(\mathcal{K}'')$ such that we can re-do the manipulations for (16) and (18), yielding $\text{Pdir}W(\mathcal{A} \cap (\mathcal{K}'' + \mathbf{x}), \mathbf{r}'', \mathbf{x})$ as a lower bound if \mathbf{r}'' is feasible from \mathbf{x} in \mathcal{K}'' . As long as we do not obtain a feasible direction, we keep re-using Lemma 5 to project the direction on a lower dimensional face of $\tilde{\mathcal{M}}$ that contains \mathbf{x} . This process must stop at some point; ultimately, we will reach the lowest dimensional face $\mathcal{K}_{\mathbf{x}}$ that contains \mathbf{x} . As \mathbf{x} lies in the relative interior of $\mathcal{K}_{\mathbf{x}}$, then all directions in $\text{span}(\mathcal{K}_{\mathbf{x}})$ are feasible, and so the projected \mathbf{r} will have to be feasible. Moreover, by stringing together the equalities of the type (16) for all the projected directions, we know that $\max_{\mathbf{e} \in \mathcal{K}_{\mathbf{x}}} \langle \mathbf{r}_{\text{final}}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle > 0$ (as we originally had $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle > 0$), and thus $\mathcal{K}_{\mathbf{x}}$ is at least one-dimensional and we also have $\mathbf{r}_{\text{final}} \neq \mathbf{0}$ (this last condition is crucial to avoid having a lower bound of zero!). This concludes the proof, and also explains why in the definition of the pyramidal width (9), we consider the pyramidal directional width for all the faces of $\text{conv}(\mathcal{A})$ and respective non-zero feasible direction \mathbf{r} . \square

C Affine Invariant Formulation

Here we provide linear convergence proofs in terms of affine invariant quantities, since all the Frank-Wolfe algorithm variants presented in this paper are affine invariant. The statements presented in the main paper above are special cases of the following more general theorems, by using the bounds (20) for the curvature constant C_f , and Theorem 6 for the affine invariant strong convexity μ_f^{A} .

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{M} by a *surjective* linear or affine map $\mathbf{A} : \tilde{\mathcal{M}} \rightarrow \mathcal{M}$, then the “old” and “new” optimization problems $\min_{\mathbf{x} \in \mathcal{M}} f(\mathbf{x})$ and $\min_{\hat{\mathbf{x}} \in \tilde{\mathcal{M}}} \hat{f}(\hat{\mathbf{x}})$ for $\hat{f}(\hat{\mathbf{x}}) := f(\mathbf{A}\hat{\mathbf{x}})$ look completely the same to the algorithm.

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 \mathbf{A} , and the Frank-Wolfe algorithm and all the variants presented here are affine invariant in the even stronger sense under arbitrary surjective \mathbf{A} [15]. (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 $\nabla \hat{f} = \mathbf{A}^T \nabla f$.)

Note however that the property of being an extremum point (vertex) of \mathcal{M} is *not* affine invariant (see [4, Section 3.1] for an example). This explains why we presented all algorithms here as working with atoms \mathcal{A} rather than vertices of the domain, thus maintaining the affine invariance of the algorithms as well as their convergence analysis.

Affine Invariant Measures of Smoothness. The affine invariant convergence analysis of the standard Frank-Wolfe algorithm by [15] crucially relies on the following measure of non-linearity of the objective function f over the domain \mathcal{M} . The (upper) *curvature constant* C_f of a convex and differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, with respect to a compact domain \mathcal{M} is defined as

$$C_f := \sup_{\substack{\mathbf{x}, \mathbf{s} \in \mathcal{M}, \gamma \in [0,1], \\ \mathbf{y} = \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle). \quad (19)$$

The definition of C_f closely mimics the fundamental descent lemma (2). The assumption of bounded curvature C_f closely corresponds to a Lipschitz assumption on the gradient of f . More precisely, if ∇f is L -Lipschitz continuous on \mathcal{M} with respect to some arbitrary chosen norm $\|\cdot\|$ in dual pairing, i.e. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L \|\mathbf{x} - \mathbf{y}\|$, then

$$C_f \leq L \operatorname{diam}_{\|\cdot\|}(\mathcal{M})^2, \quad (20)$$

where $\operatorname{diam}_{\|\cdot\|}(\cdot)$ denotes the $\|\cdot\|$ -diameter, see [15, Lemma 7]. While the early papers [9, 8] 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. The quantity C_f combines the complexity of the domain \mathcal{M} and the curvature of the objective function f into a single quantity. The advantage of this combination is well illustrated in [21, Lemma A.1], where Frank-Wolfe was used to optimize a quadratic function over product of probability simplices with an exponential number of dimensions. In this case, the Lipschitz constant could be exponentially worse than the curvature constant which does take the simplex geometry of \mathcal{M} into account.

An Affine Invariant Notion of Strong Convexity which Depends on the Geometry of \mathcal{M} . We now present the affine invariant analog of the strong convexity bound (4), which could be interpreted as the *lower* curvature μ_f^A analog of C_f . The role of γ in the definition (19) of C_f was to define an affine invariant scale by only looking at proportions over lines (as line segments between \mathbf{x} and \mathbf{s} in this case). The trick here is to use anchor points in \mathcal{A} 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 FW atom and the away atom appearing in the away-steps FW algorithm. Specifically, let \mathbf{x}^* be a potential optimal point and \mathbf{x} a non-optimal point; thus we have $\langle -\nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle > 0$ (i.e. $\mathbf{x}^* - \mathbf{x}$ is a strict descent direction from \mathbf{x} for f). We then define the positive step-size quantity:

$$\gamma^A(\mathbf{x}, \mathbf{x}^*) := \frac{\langle -\nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle}{\langle -\nabla f(\mathbf{x}), \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}. \quad (21)$$

This quantity is motivated from both (6) and the linear rate inequality (5), and enables to transfer lengths from the error $\mathbf{e}_t = \mathbf{x}^* - \mathbf{x}_t$ to the pairwise FW direction $\mathbf{d}_t^{\text{PFW}} = \mathbf{s}_f(\mathbf{x}_t) - \mathbf{v}_f(\mathbf{x}_t)$. More precisely, $\mathbf{s}_f(\mathbf{x}) := \arg \min_{\mathbf{v} \in \mathcal{A}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ is the standard FW atom. To define the away-atom, we consider all possible expansions of \mathbf{x} as a convex combination of atoms.¹⁰ We recall that set of possible active sets is $\mathcal{S}_\mathbf{x} := \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{A} \text{ such that } \mathbf{x} \text{ is a proper convex combination of all the elements in } \mathcal{S}\}$. For a given set \mathcal{S} , we write $\mathbf{v}_\mathcal{S}(\mathbf{x}) := \arg \max_{\mathbf{v} \in \mathcal{S}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ for the away atom in the algorithm supposing that the current set of active atoms is \mathcal{S} . Finally, we define $\mathbf{v}_f(\mathbf{x}) := \arg \min_{\{\mathbf{v} = \mathbf{v}_\mathcal{S}(\mathbf{x}) \mid \mathcal{S} \in \mathcal{S}_\mathbf{x}\}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ to be the worst-case away atom (that is, the atom which would yield the smallest away descent).

We then define the *geometric strong convexity* constant μ_f^A which depends *both* on the function f and the domain $\mathcal{M} = \operatorname{conv}(\mathcal{A})$:

$$\mu_f^A := \inf_{\mathbf{x} \in \mathcal{M}} \inf_{\substack{\mathbf{x}^* \in \mathcal{M} \\ \text{s.t. } \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle < 0}} \frac{2}{\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle). \quad (22)$$

C.1 Lower Bound for the Geometric Strong Convexity Constant μ_f^A

The geometric strong convexity constant μ_f^A , as defined in (22), 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{M} . Theorem 6 allows us to lower bound the constant μ_f^A in terms of the strong convexity of the objective function, combined with a purely geometric complexity measure of the domain \mathcal{M} (its pyramidal width $PWidth(\mathcal{A})$ (9)). In the following Section D below, we will show the linear convergence of the four variants of the FW algorithm presented in this paper under the assumption that $\mu_f^A > 0$.

¹⁰As we are working with general polytopes, the expansion of a point as a convex combination of atoms is not necessarily unique.

In view of the following Theorem 6, we have that the condition $\mu_f^A > 0$ is slightly weaker than the strong convexity of the objective function¹¹ over a polytope domain (it is implied by strong convexity).

Theorem 6. *Let f be a convex differentiable function and suppose that f is μ -strongly convex w.r.t. to the Euclidean norm $\|\cdot\|$ over the domain $\mathcal{M} = \text{conv}(\mathcal{A})$ with strong-convexity constant $\mu \geq 0$. Then*

$$\mu_f^A \geq \mu \cdot (\text{PWidth}(\mathcal{A}))^2. \quad (23)$$

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

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (24)$$

Using the strong convexity bound (24) with $\mathbf{y} := \mathbf{x}^*$ on the right hand side of equation (22) (and using the shorthand $\mathbf{r}_\mathbf{x} := -\nabla f(\mathbf{x})$), we thus get:

$$\begin{aligned} \mu_f^A &\geq \inf_{\substack{\mathbf{x}, \mathbf{x}^* \in \mathcal{M} \\ \text{s.t. } \langle \mathbf{r}_\mathbf{x}, \mathbf{x}^* - \mathbf{x} \rangle > 0}} \mu \left(\frac{\langle \mathbf{r}_\mathbf{x}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_\mathbf{x}, \mathbf{x}^* - \mathbf{x} \rangle} \|\mathbf{x}^* - \mathbf{x}\| \right)^2 \\ &= \mu \inf_{\substack{\mathbf{x} \neq \mathbf{x}^* \in \mathcal{M} \\ \text{s.t. } \langle \mathbf{r}_\mathbf{x}, \hat{\mathbf{e}}(\mathbf{x}^*, \mathbf{x}) \rangle > 0}} \left(\frac{\langle \mathbf{r}_\mathbf{x}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_\mathbf{x}, \hat{\mathbf{e}}(\mathbf{x}^*, \mathbf{x}) \rangle} \right)^2, \end{aligned} \quad (25)$$

where $\hat{\mathbf{e}}(\mathbf{x}^*, \mathbf{x}) := \frac{\mathbf{x}^* - \mathbf{x}}{\|\mathbf{x}^* - \mathbf{x}\|}$ is the unit length 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{M}) with the additional constraint that it makes a positive inner product with the negative gradient $\mathbf{r}_\mathbf{x}$, i.e. it is a strict descent direction. This is only possible if \mathbf{x} is not already optimal, i.e. $\mathbf{x} \in \mathcal{M} \setminus \mathcal{X}^*$ where $\mathcal{X}^* := \{\mathbf{x}^* \in \mathcal{M} : \langle \mathbf{r}_{\mathbf{x}^*}, \mathbf{x} - \mathbf{x}^* \rangle \leq 0 \ \forall \mathbf{x} \in \mathcal{M}\}$ is the set of optimal points.

We note that $\mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x})$ is a valid pairwise FW direction for a specific active set \mathcal{S} for \mathbf{x} , and so we can re-use (12) from Theorem 3 for the right hand side of (25) to conclude the proof. \square

We now proceed to present the main linear convergence result in the next section, using only the mentioned affine invariant quantities.

D Linear Convergence Proofs

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:

$$C_f^A := \sup_{\substack{\mathbf{x}, \mathbf{s}, \mathbf{v} \in \mathcal{M}, \\ \gamma \in [0, 1], \\ \mathbf{y} = \mathbf{x} + \gamma(\mathbf{s} - \mathbf{v})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \gamma \langle \nabla f(\mathbf{x}), \mathbf{s} - \mathbf{v} \rangle). \quad (26)$$

By comparing with C_f (19), we see that the modification is that \mathbf{y} is defined with any direction $\mathbf{s} - \mathbf{v}$ instead of a standard FW direction $\mathbf{s} - \mathbf{x}$. This allows to use the away direction or the pairwise FW direction even though these might yield some \mathbf{y} 's which are outside of the domain \mathcal{M} when using $\gamma > \gamma_{\max}$ (in fact, $\mathbf{y} \in \mathcal{M}^A := \mathcal{M} + (\mathcal{M} - \mathcal{M})$ in the Minkowski sense). On the other hand, by re-using a similar argument as in [15, Lemma 7], we can obtain the same bound (20) for C_f^A , with the only difference that the Lipschitz constant L for the gradient function has to be valid on \mathcal{M}^A instead of just \mathcal{M} .

Remark 7. *For all pairs of functions f and compact domains \mathcal{M} , it holds that $\mu_f^A \leq C_f$ (and $C_f \leq C_f^A$).*

¹¹As an example of function that is not strongly convex but can still have $\mu_f^A > 0$, consider $f(\mathbf{x}) := g(\mathbf{A}\mathbf{x})$ where g is μ_g -strongly convex, but the matrix \mathbf{A} is rank deficient. Then by using the affine invariance of the definition of μ_f^A and using Theorem (6) applied on the equivalent problem on g with domain $\text{conv}(\mathbf{A}\mathcal{A})$, we get $\mu_f^A \geq \mu_g \cdot (\text{PWidth}(\mathbf{A}\mathcal{A}))^2 > 0$.

Proof. Let \mathbf{x} be a vertex of \mathcal{M} , so that $\mathcal{S} = \{\mathbf{x}\}$. Then $\mathbf{x} = \mathbf{v}_f(\mathbf{x})$. Pick $\mathbf{x}^* := \mathbf{s}_f(\mathbf{x})$ and substitute in the definition for μ_f^A (22). 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 (19). Thus, we have $\mu_f^A \leq 2(f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle) \leq C_f$. \square

We now give the global linear convergence rates for the four variants of the FW algorithm: away-steps FW (AFW Algorithm 1); pairwise FW (PFW Algorithm 2); fully-corrective FW (FCFW Algorithm 3 with approximate correction as per Algorithm 4); and Wolfe’s min-norm point algorithm (Algorithm 3 with MNP-correction given in Algorithm 5). For the AFW, MNP and PFW algorithms, we call a *drop step* when the active set shrinks, i.e. $|S^{(t+1)}| < |S^{(t)}|$. For the PFW algorithm, we also have the possibility of a *swap step*, where $\gamma_t = \gamma_{\max}$ but the size of the active set stays constant $|S^{(t+1)}| = |S^{(t)}|$ (i.e. the mass gets fully swapped from the away atom to the FW atom). We note that a nice property of the FCFW variant is that it does not have any drop steps (it executes both FW steps and away steps simultaneously while guaranteeing enough progress at every iteration).

Theorem 8. Suppose that f has smoothness constant C_f^A (C_f for FCFW and MNP), as well as geometric strong convexity constant μ_f^A as defined in (22). Then the suboptimality $h_t := f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)$ of the iterates of all the four variants of the FW algorithm decreases geometrically at each step that is not a drop step nor a swap step (i.e. when $\gamma_t < \gamma_{\max}$, called a ‘good step’¹²), that is

$$h_{t+1} \leq (1 - \rho_f) h_t,$$

where:

$$\begin{aligned} \rho_f &:= \frac{\mu_f^A}{4C_f^A} \quad \text{for the AFW algorithm,} & \rho_f &:= \min \left\{ \frac{1}{2}, \frac{\mu_f^A}{C_f^A} \right\} \quad \text{for the PFW algorithm,} \\ \rho_f &:= \frac{\mu_f^A}{4C_f} \quad \text{for the FCFW algorithm,} & \rho_f &:= \min \left\{ \frac{1}{2}, \frac{\mu_f^A}{C_f} \right\} \quad \text{for the MNP algorithm, or} \\ & & & \text{FCFW with exact correction.} \end{aligned}$$

Moreover, the number of drop steps up to iteration t is bounded by $t/2$. This yields the global linear convergence rate of $h_t \leq h_0 \exp(-\frac{1}{2}\rho_f t)$ for the AFW and MNP variants. FCFW does not need the extra $1/2$ factor as it does not have any bad step. Finally, the PFW algorithm has at most $3|\mathcal{A}|!$ swap steps between any two ‘good steps’.

If $\mu_f^A = 0$ (i.e. the case of general convex objectives), then all the four variants have a $O(1/k(t))$ convergence rate where $k(t)$ is the number of ‘good steps’ up to iteration t . More specifically, we can summarize the suboptimality bounds for the four variants as:

$$h_t \leq \frac{4C}{k(t) + 4} \quad \text{for } k(t) \geq 1,$$

where $C = 2\mu_f^A + h_0$ for AFW; $C = 2C_f + h_0$ for FCFW with approximate correction; $C = C_f/2$ for MNP; and $C = C_f^A/2$ for PFW. The number of good steps is $k(t) = t$ for FCFW; it is $k(t) \geq t/2$ for MNP and AFW; and $k(t) \geq t/(3|\mathcal{A}|! + 1)$ for PFW.

Proof. Proof for AFW. The general idea of the proof is to use the definition of the geometric strong convexity constant to upper bound h_t , while using the definition of the curvature constant C_f^A to lower bound the decrease in primal suboptimality $h_t - h_{t+1}$ for the ‘good steps’ of Algorithm 1. Then we upper bound the number of ‘bad steps’ (the drop steps).

Upper bounding h_t . In the whole proof, we assume that $\mathbf{x}^{(t)}$ is not already optimal, i.e. that $h_t > 0$. If $h_t = 0$, then because line-search is used, we will have $h_{t+1} \leq h_t = 0$ and so the geometric rate of decrease is trivially true in this case. Let \mathbf{x}^* be an optimum point (which is not necessarily unique). As $h_t > 0$, we have that $\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{x}^* - \mathbf{x}^{(t)} \rangle > 0$. We can thus apply the geometric strong

¹²Note that any step with $\gamma_{\max} \geq 1$ can also be considered a ‘good step’, even if $\gamma_t = \gamma_{\max}$, as is apparent from the proof. The problematic steps arise only when $\gamma_{\max} \ll 1$.

convexity bound (22) at the current iterate $\mathbf{x} := \mathbf{x}^{(t)}$ using \mathbf{x}^* as an optimum reference point to get (with $\bar{\gamma} := \gamma^A(\mathbf{x}^{(t)}, \mathbf{x}^*)$ as defined in (21)):

$$\begin{aligned} \frac{\bar{\gamma}^2}{2} \mu_f^A &\leq f(\mathbf{x}^*) - f(\mathbf{x}^{(t)}) + \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{x}^* - \mathbf{x}^{(t)} \right\rangle \\ &= -h_t + \bar{\gamma} \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{s}_f(\mathbf{x}^{(t)}) - \mathbf{v}_f(\mathbf{x}^{(t)}) \right\rangle \\ &\leq -h_t + \bar{\gamma} \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{s}_t - \mathbf{v}_t \right\rangle \\ &= -h_t + \bar{\gamma} g_t, \end{aligned} \quad (27)$$

where we define $g_t := \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{s}_t - \mathbf{v}_t \right\rangle$ (note that $h_t \leq g_t$ and so g_t also gives a primal suboptimality certificate). For the third line, we have used the definition of $\mathbf{v}_f(\mathbf{x})$ which implies $\left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{v}_f(\mathbf{x}^{(t)}) \right\rangle \leq \left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{v}_t \right\rangle$. Therefore $h_t \leq -\frac{\bar{\gamma}^2}{2} \mu_f^A + \bar{\gamma} g_t$, which is always upper bounded¹³ by

$$h_t \leq \frac{g_t^2}{2\mu_f^A}. \quad (28)$$

Lower bounding progress $h_t - h_{t+1}$. We here use the key aspect in the proof that we had described in the main text with (6). Because of the way the direction \mathbf{d}_t is chosen in the AFW Algorithm 1, we have

$$\left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t \right\rangle \geq g_t/2, \quad (29)$$

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

We first consider the case $\gamma_{\max} \geq 1$. Let $\mathbf{x}_\gamma := \mathbf{x}^{(t)} + \gamma \mathbf{d}_t$ be the point obtained by moving with step-size γ in direction \mathbf{d}_t , where \mathbf{d}_t is the one chosen by Algorithm 1. By using $\mathbf{s} := \mathbf{x}^{(t)} + \mathbf{d}_t$ (a feasible point as $\gamma_{\max} \geq 1$), $\mathbf{x} := \mathbf{x}^{(t)}$ and $\mathbf{y} := \mathbf{x}_\gamma$ in the definition of the curvature constant C_f (19), and solving for $f(\mathbf{x}_\gamma)$, we get the affine invariant version of the descent lemma (2):

$$f(\mathbf{x}_\gamma) \leq f(\mathbf{x}^{(t)}) + \gamma \left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t \right\rangle + \frac{\gamma^2}{2} C_f, \quad \text{valid } \forall \gamma \in [0, 1]. \quad (30)$$

As γ_t is obtained by line-search and that $[0, 1] \subseteq [0, \gamma_{\max}]$, we also have that $f(\mathbf{x}^{(t+1)}) = f(\mathbf{x}_{\gamma_t}) \leq f(\mathbf{x}_\gamma) \forall \gamma \in [0, 1]$. Combining these two inequalities, subtracting $f(\mathbf{x}^*)$ on both sides, and using $C_f \leq C_f^A$ to simplify the possibilities yields $h_{t+1} \leq h_t + \gamma \left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t \right\rangle + \frac{\gamma^2}{2} C_f^A$.

Using the crucial gap inequality (29), we get $h_{t+1} \leq h_t - \gamma \frac{g_t}{2} + \frac{\gamma^2}{2} C_f^A$, and so:

$$h_t - h_{t+1} \geq \gamma \frac{g_t}{2} - \frac{\gamma^2}{2} C_f^A \quad \forall \gamma \in [0, 1]. \quad (31)$$

We can minimize the bound (31) on the right hand side by letting $\gamma = \gamma_t^B := \frac{g_t}{2C_f^A}$. Supposing that $\gamma_t^B \leq 1$, we then get $h_t - h_{t+1} \geq \frac{g_t^2}{8C_f^A}$ (we cover the case $\gamma_t^B > 1$ later). By combining this inequality with the one from geometric strong convexity (28), we get

$$h_t - h_{t+1} \geq \frac{\mu_f^A}{4C_f^A} h_t \quad (32)$$

implying that we have a geometric rate of decrease $h_{t+1} \leq \left(1 - \frac{\mu_f^A}{4C_f^A}\right) h_t$ (this is a ‘good step’).

Boundary cases. We now consider the case $\gamma_t^B > 1$ (with $\gamma_{\max} \geq 1$ still). The condition $\gamma_t^B > 1$ then translates to $g_t \geq 2C_f^A$, which we can use in (31) with $\gamma = 1$ to get $h_t - h_{t+1} \geq \frac{g_t}{2} - \frac{g_t}{4} = \frac{g_t}{4}$.

¹³Here we have used the trivial inequality $0 \leq a^2 - 2ab + b^2$ for the choice of numbers $a := \frac{g_t}{\mu_f^A}$ and $b := \bar{\gamma}$.

An alternative way to obtain the bound is to look at the unconstrained maximum of the RHS which is a concave function of $\bar{\gamma}$ by letting $\bar{\gamma} = g_t/\mu_f^A$, as we did in the main paper to obtain the upper bound on h_t in (5).

Combining this inequality with $h_t \leq g_t$ gives the geometric decrease $h_{t+1} \leq (1 - \frac{1}{4}) h_t$ (also a ‘good step’). ρ_f^A is obtained by considering the worst-case of the constants obtained from $\gamma_t^B > 1$ and $\gamma_t^B \leq 1$. (Note that $\mu_f^A \leq C_f^A$ by Remark 7, and thus $\frac{1}{4} \geq \frac{\mu_f^A}{4C_f^A}$).

Finally, we are left with the case that $\gamma_{\max} < 1$. This is thus an away step and so $\mathbf{d}_t = \mathbf{d}_t^A = \mathbf{x}^{(t)} - \mathbf{v}_t$. Here, we use the away version C_f^A : by letting $\mathbf{s} := \mathbf{x}^{(t)}$, $\mathbf{v} = \mathbf{v}_t$ and $\mathbf{y} := \mathbf{x}_\gamma$ in (26), we also get the bound $f(\mathbf{x}_\gamma) \leq f(\mathbf{x}^{(t)}) + \gamma \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{d}_t \rangle + \frac{\gamma^2}{2} C_f^A$, valid $\forall \gamma \in [0, 1]$ (but note here that the points \mathbf{x}_γ are not feasible for $\gamma > \gamma_{\max}$ – the bound considers some points outside of \mathcal{M}). We now have two options: either $\gamma_t = \gamma_{\max}$ (a drop step) or $\gamma_t < \gamma_{\max}$. In the case $\gamma_t < \gamma_{\max}$ (the line-search yields a solution in the interior of $[0, \gamma_{\max}]$), then because $f(\mathbf{x}_\gamma)$ is convex in γ , we know that $\min_{\gamma \in [0, \gamma_{\max}]} f(\mathbf{x}_\gamma) = \min_{\gamma \geq 0} f(\mathbf{x}_\gamma)$ and thus $\min_{\gamma \in [0, \gamma_{\max}]} f(\mathbf{x}_\gamma) = f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}_\gamma) \forall \gamma \in [0, 1]$. We can then re-use the same argument above equation (31) to get the inequality (31), and again considering both the case $\gamma_t^B \leq 1$ (which yields inequality (32)) and the case $\gamma_t^B > 1$ (which yields $(1 - \frac{1}{4})$ as the geometric rate constant), we get a ‘good step’ with $1 - \rho_f$ as the worst-case geometric rate constant.

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

Proof for FCFW. In the case of FCFW, we do not need to consider away steps: by the quality of the approximate correction in Algorithm 4 (as specified in Line 4), we know that at the beginning of a new iteration, the away gap $g_t^A \leq \epsilon$. Supposing that the algorithm does not exit at line 6 of Algorithm 3, then $g_t^{\text{FW}} > \epsilon$ and thus we have that $2\langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} \rangle \geq \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle$ using a similar argument as in (6) (i.e. if one would be to run the AFW algorithm at this point, it would take a FW step). Finally, by property of the line 3 of the approximate correction Algorithm 4, the correction is guaranteed to make at least as much progress as a line-search in direction \mathbf{d}_t^{FW} , and so the lower bound (31) can be used for FCFW as well (but using C_f as the constant instead of C_f^A given that it was a FW step).

Proof for MNP. After a correction step in the MNP algorithm, we have that the current iterate is the minimizer over the active set, and thus $g_t^A = 0$. We thus have $\langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} \rangle = \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle = g_t$, which means that a standard FW step would yield a geometric decrease of error.¹⁴ It thus remains to show that the MNP-correction is making as much progress as a FW line-search. Consider \mathbf{y}_1 as defined in Algorithm 5. If it belongs to $\text{conv}(\mathcal{V}^{(0)})$, then it has made more progress than a FW line-search as \mathbf{s}_t and $\mathbf{x}^{(t)}$ belongs to $\text{conv}(\mathcal{V}^{(0)})$.

The next possibility is the crucial step in the proof: suppose that exactly one atom was removed from the correction polytope and that \mathbf{y}_1 does not belong to $\text{conv}(\mathcal{V}^{(0)})$ (as this was covered in the above case). This means that \mathbf{y}_2 belongs to the *relative interior* of $\text{conv}(\mathcal{V}^{(1)})$. Because \mathbf{y}_2 is by definition the affine minimizer of f on $\text{conv}(\mathcal{V}^{(1)})$, the negative gradient $-\nabla f(\mathbf{y}_2)$ is pointing away to the polytope $\text{conv}(\mathcal{V}^{(1)})$ (by the optimality condition). But $\text{conv}(\mathcal{V}^{(1)})$ is a *facet* of $\text{conv}(\mathcal{V}^{(0)})$, this means that $-\nabla f(\mathbf{y}_2)$ determines a facet of $\text{conv}(\mathcal{V}^{(0)})$ (i.e. $\langle -\nabla f(\mathbf{y}_2), \mathbf{y} - \mathbf{y}_2 \rangle \leq 0$ for all $\mathbf{y} \in \text{conv}(\mathcal{V}^{(0)})$). This means that \mathbf{y}_2 is also the minimizer of f on $\text{conv}(\mathcal{V}^{(0)})$ and thus has made more progress than a FW line-search.

In the case that two atoms are removed from $\text{conv}(\mathcal{V}^{(0)})$, we cannot make this argument anymore (it is possible that \mathbf{y}_3 makes less progress than a FW line-search); but in this case, the size of the active set is reduced by one (we have a drop step), and thus we can use the same argument as in the AFW algorithm to bound the number of such steps.

¹⁴Moreover, as we do not have the factor of 2 relating $\langle \mathbf{r}_t, \mathbf{d}_t^{\text{FW}} \rangle$ and g_t unlike in the AFW and approximate FCFW case, we can remove the factor of $\frac{1}{2}$ in front of g_t in (31), removing the factor of $\frac{1}{4}$ appearing in (32), and also giving a geometric decrease with factor $(1 - \frac{1}{2})$ when $\gamma_t^B > 1$.

Proof for PFW. In this case, $\langle \mathbf{r}_t, \mathbf{d}_t \rangle = \langle \mathbf{r}_t, \mathbf{d}_t^{\text{PFW}} \rangle$, so we do not even need a factor of 2 to relate the gaps (with the same consequence as in MNP in getting slightly bigger constants). We can re-use the same argument as in the AFW algorithm to get a geometric progress when $\gamma_t < \gamma_{\max}$. When $\gamma_t = \gamma_{\max}$ we can either have a drop step if \mathbf{s}_t was already in $\mathcal{S}^{(t)}$, or a swap step if \mathbf{s}_t was also added to $\mathcal{S}^{(t)}$ and so $|\mathcal{S}^{(t+1)}| = |\mathcal{S}^{(t)}|$. The number of drop steps can be bounded similarly as in the AFW algorithm. On the other hand, in the worst case, there could be a very large number of swap steps. We provide here a very loose bound, though it would be interesting to use other properties of the objective to prove that this worst case scenario cannot happen.

We thus bound the maximum number of swap steps between two ‘good steps’ (very loosely). Let $m = |\mathcal{A}|$ be the number of possible atoms, and let r be the size of the current active set $|\mathcal{S}^{(t)}| = r \leq m$. When doing a drop step $\gamma_t = \alpha_{v_t}$, there are two possibilities: either we move all the mass from v_t to a new atom $\mathbf{s}_t \notin \mathcal{S}^{(t)}$ i.e. $\alpha_{v_t}^{(t+1)} = 0$ and $\alpha_{\mathbf{s}_t}^{(t+1)} = \alpha_{v_t}^{(t)}$ (a swap step); or we move all the mass from v_t to an old atom $\mathbf{s}_t \in \mathcal{S}^{(t)}$ i.e. $\alpha_{\mathbf{s}_t}^{(t+1)} = \alpha_{\mathbf{s}_t}^{(t)} + \alpha_{v_t}^{(t)}$ (a ‘full drop step’). When doing a swap step, the set of *possible values* for the coordinates α_v *do not change*, they are only ‘swapped around’ amongst the m possible slots. The maximum number of possible consecutive swap steps without revisiting an iterate already seen is thus bounded by the number of ways we can assign r numbers in m slots (supposing the r coordinates were all distinct in the worst case), which is $m!/(m-r)!$. Note that because the algorithm does a line-search in a strict descent direction at each iteration, we always have $f(\mathbf{x}^{(t+1)}) < f(\mathbf{x}^{(t)})$ unless $\mathbf{x}^{(t)}$ is already optimal. This means that the algorithm cannot revisit the same point unless it has converged. When doing a ‘full drop step’, the set of coordinates changes, but the size of the active set is reduced by one (thus r reduced by one). In the worst case, we will do a maximum number of swap steps, followed by a full drop step, repeated so on all the way until we reach an active set of only one element (in which case there is a maximum number of m swap steps). Starting with an active set of r coordinates, the maximum number of swap steps B without doing any ‘good step’ (which would also change the set of coordinates), is thus upper bounded by:

$$B \leq \sum_{l=1}^r \frac{m!}{(m-l)!} \leq m! \sum_{l=0}^{\infty} \frac{1}{l!} = m!e \leq 3m!,$$

as claimed.

Proof of Sublinear Convergence for General Convex Objectives (i.e. when $\mu_f^A = 0$). For the good steps of the MNP algorithm and the pairwise FW algorithm, we have the reduction of suboptimality given by (31) without the factor of $\frac{1}{2}$ in front of $g_t \geq h_t$. This is the standard recurrence that appears in the convergence proof of Frank-Wolfe (see for example Equation (4) in [15, proof of Theorem 1]), yielding the usual convergence:

$$h_t \leq \frac{2C}{k(t) + 2} \quad \text{for } k(t) \geq 1, \quad (33)$$

where $k(t)$ is the number of good steps up to iteration t , and $C = C_f$ for MNP and $C = C_f^A$ for PFW. The number of good steps for MNP is $k(t) \geq t/2$, while for PFW, we have the (useless) lower bound $k(t) \geq t/(3|\mathcal{A}| + 1)$. For FCFW with exact correction, the rate (33) was already proven in [15] with $k(t) = t$. On the other hand, for FCFW with approximate correction, and for AFW, the factor of $\frac{1}{2}$ in front of the gap g_t in the suboptimality bound (31) somewhat complicates the convergence proof. The recurrence we get for the suboptimality is the same as in Equation (20) of [21, proof of Theorem C.1], with $\nu = \frac{1}{2}$ and $n = 1$, giving the following suboptimality bound:

$$h_t \leq \frac{4C}{k(t) + 4} \quad \text{for } k(t) \geq 0, \quad (34)$$

where $C = 2C_f^A + h_0$ for AFW and $C = 2C_f + h_0$ for FCFW with approximate correction. Moreover, the number of good steps is $k(t) \geq t/2$ for AFW, and $k(t) = t$ for FCFW. A weaker (logarithmic) dependence on the initial error h_0 can also be obtained by following a tighter analysis (see [21, Theorem C.4] or [18, Lemma D.5 and Theorem D.6]), though we only state the simpler result here. \square

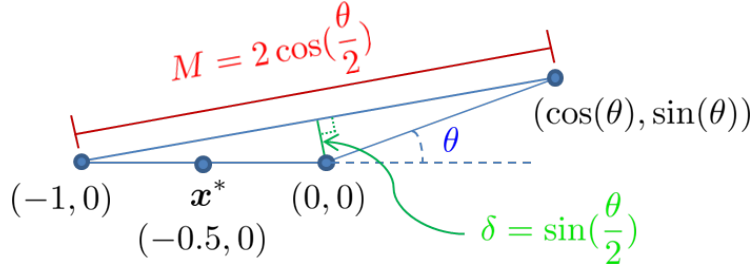


Figure 4: Simple triangle domain to test the empirical tightness of the constant in the convergence rate for AFW and PFW. The width δ varies with θ . We optimize $f(x) := \frac{1}{2}\|x - x^*\|^2$ over this domain.

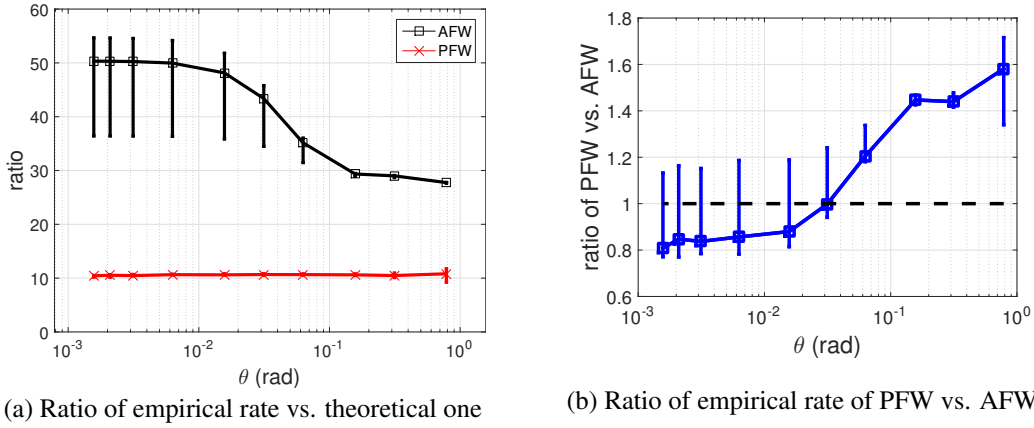


Figure 5: Empirical rate results for the triangle domain of Figure 4. We plot median values over 20 random starting points; the error bars represent the 25% and 75% quantiles. The empirical rate for PFW is closely following the theoretical one.

E Empirical Tightness of Linear Rate Constant

We describe here a simple experiment to test how tight the constant in the linear convergence rate of Theorem 8 is. We test both AFW and PFW on the triangle domain with corners at the locations $(-1, 0)$, $(0, 0)$ and $(\cos(\theta), \sin(\theta))$, for increasingly small θ (see Figure 4).

The pyramidal width $\delta = \sin(\frac{\theta}{2})$ becomes vanishingly small as $\theta \rightarrow 0$; the diameter is $M = 2 \cos(\frac{\theta}{2})$. We consider the optimization of the function $f(x) := \frac{1}{2}\|x - x^*\|^2$ with $x^* = (-0.5, 1)$ on one edge of the domain. Note that the condition number of f is $\frac{L}{\mu} = 1$. The bound on the linear convergence rate ρ according to Theorem 8 (using $C_f^A \leq LM^2$ (20) and $\mu_f^A \geq \mu \delta^2$ (23)) is $\rho^{\text{PFW}} = \frac{\mu_f^A}{C_f^A} \geq \frac{\mu}{L} \left(\frac{\delta}{M}\right)^2$ for PFW and $\rho^A = \frac{1}{4}\rho^{\text{PFW}}$ for AFW. The theoretical constant here is thus $\rho^{\text{PFW}} = \frac{1}{4} \tan^2(\frac{\theta}{2})$. We consider θ varying from $\pi/4$ to $1e-3$, and thus theoretical rates varying on a wide range from 0.04 to $1e-7$. We compare the theoretical rate ρ with the empirically observed one by estimating $\hat{\rho}$ in the relationship $h_t \approx h_0 \exp(-\rho t)$ (using linear regression on the semilogarithmic scale). For each θ , we run both AFW and PFW for 2000 iterations starting from 20 different random starting points¹⁵ in the interior of the triangle domain. We disregard the starting points that yield a drop step (as then the algorithm converges in one iteration; these happen for about 10% of the starting points). Note that as there is no drop step in our setup, we do not need to divide by two the effective rate as is done in Theorem 8 (the number of ‘good steps’ is $k(t) = t$).

¹⁵ x_0 is obtained by taking a random convex combination of the corners of the domain.

Figure 5 presents the results. In Figure 5(a), we plot the ratio of the estimated rate over the theoretical rate $\frac{\hat{\rho}}{\rho}$ for both PFW and AFW as θ varies. Note that the ratio is very stable for PFW (around 10), despite the rate changing through six orders of magnitude, demonstrating the empirical tightness of the constant for this domain. The ratio for AFW has more fluctuations, but also stays within a stable range. We can also do a finer analysis than the pyramidal width and consider the finite number possibilities for the worst case angles for $(\langle \hat{\mathbf{r}}, \mathbf{d}^{\text{PFW}}(\mathbf{r}) \rangle)^2$. This gives the tighter constant $\rho^{\text{PFW}} = \sin^2(\frac{\theta}{2})$ for our triangle domain, gaining a factor of about 4, but still not matching yet the empirical observation for PFW.

In Figure 5(b), we compare the empirical rate for PFW vs. the one for AFW. For bigger theoretical rates, PFW appears to converge faster. However, AFW gets a slightly better empirical rate for very small rates (small angles).

F Non-Strongly Convex Generalization

Here we will study the generalized setting with objective $f(\mathbf{x}) := g(\mathbf{A}\mathbf{x}) + \langle \mathbf{b}, \mathbf{x} \rangle$ where g is μ_g -strongly convex w.r.t. the Euclidean norm over the domain \mathcal{AM} with strong convexity constant $\mu_g > 0$.

We first define a few constants: let $G := \max_{\mathbf{x} \in \mathcal{M}} \|\nabla g(\mathbf{A}\mathbf{x})\|$ be the maximal norm of the gradient of g over \mathcal{AM} ; M be the diameter of \mathcal{M} and $M_{\mathbf{A}}$ be the diameter of \mathcal{AM} .

Let θ be the Hoffman constant (see [4, Lemma 2.2]) associated with the matrix $[\mathbf{A}; \mathbf{b}^\top; \mathbf{B}] = \begin{pmatrix} \mathbf{A} \\ \mathbf{b}^\top \\ \mathbf{B} \end{pmatrix}$, where the rows of \mathbf{B} are the linear inequality constraints defining the set \mathcal{M} .

We present here a generalization of Lemma 2.5 from [4]:

Lemma 9. *For any $\mathbf{x} \in \mathcal{M}$ and \mathbf{x}^* in the solution set \mathcal{X}^* :*

$$f(\mathbf{x}^*) - f(\mathbf{x}) - 2\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \geq 2\tilde{\mu} d(\mathbf{x}, \mathcal{X}^*)^2, \quad (35)$$

where $\tilde{\mu} := 1 / \left(2\theta^2 \left(\|\mathbf{b}\|M + 3GM_{\mathbf{A}} + \frac{2}{\mu_g}(G^2 + 1) \right) \right)$ is the generalized strong convexity constant for f .

Proof. Let \mathbf{x}^* be any element of the solution set \mathcal{X}^* . By the strong convexity of g , we have

$$f(\mathbf{x}^*) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \geq \frac{\mu_g}{2} \|\mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}\|^2. \quad (36)$$

Moreover, by the convexity of f , we have:

$$-\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \geq f(\mathbf{x}) - f(\mathbf{x}^*). \quad (37)$$

We now use inequality (2.10) in [4] to get the bound:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{1}{B_1} (\langle \mathbf{b}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x}^* \rangle)^2, \quad (38)$$

where $B_1 := (\|\mathbf{b}\|M + 3GM_{\mathbf{A}} + \frac{2}{\mu_g}G^2)$.

Plugging (38) into (37) and adding to (36), we get:

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}) - 2\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle &\geq \frac{1}{B_2} \left(\|\mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}\|^2 + (\langle \mathbf{b}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x}^* \rangle)^2 \right) \\ &\geq \frac{1}{B_2\theta^2} d(\mathbf{x}, \mathcal{X}^*)^2, \end{aligned}$$

where $B_2 := (\|\mathbf{b}\|M + 3GM_{\mathbf{A}} + \frac{2}{\mu_g}(G^2 + 1))$. For the last inequality, we used inequality (2.1) in [4] that made use of the Hoffman's Lemma (see [4, Lemma 2.2]), where θ is the Hoffman constant associated with the matrix $[\mathbf{A}; \mathbf{b}^\top; \mathbf{B}]$. In this case, \mathbf{B} is the matrix with rows containing the linear inequality constraints defining \mathcal{M} . \square

We now define the following generalization of the geometric strong convexity constant (22), that we now call $\tilde{\mu}_f$:

$$\tilde{\mu}_f := \inf_{\mathbf{x} \in \mathcal{M}} \sup_{\substack{\mathbf{x}^* \in \mathcal{X}^* \\ \text{s.t. } \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle < 0}} \frac{1}{2\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}) - 2 \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle). \quad (39)$$

Notice the new inner *supremum* over the solution set \mathcal{X}^* compared to the original definition (22), the factor of 2 in front of the gradient, and the different overall scaling to have a similar form as in the previous linear convergence theorem. This new quantity $\tilde{\mu}_f$ is still *affine invariant*, but unfortunately now depends on the location of the solution set \mathcal{X}^* . We now present the generalization of Theorem 6.

Theorem 10. *Let $f(\mathbf{x}) := g(\mathbf{A}\mathbf{x}) + \langle \mathbf{b}, \mathbf{x} \rangle$ where g is μ_g -strongly convex w.r.t. the Euclidean norm over the domain \mathcal{AM} with strong convexity constant $\mu_g > 0$. Let $\tilde{\mu}$ be the corresponding generalized strong convexity constant coming from Lemma 9. Then*

$$\tilde{\mu}_f \geq \tilde{\mu} \cdot (\text{PWidth}(\mathcal{M}))^2.$$

Proof. Let \mathbf{x} be fixed and not optimal; let \mathbf{x}^* be its closest point in \mathcal{X}^* i.e. $\|\mathbf{x} - \mathbf{x}^*\| = d(\mathbf{x}, \mathcal{X}^*)$. We have that $\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle < 0$ as \mathbf{x} is not optimal.

We use the generalized strong convexity notion $\tilde{\mu}$ from Lemma 9 for the particular reference point \mathbf{x}^* in the third line below to get:

$$\begin{aligned} & \sup_{\substack{\mathbf{x}' \in \mathcal{X}^* \\ \text{s.t. } \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle < 0}} \frac{1}{2\gamma^A(\mathbf{x}, \mathbf{x}')^2} (f(\mathbf{x}') - f(\mathbf{x}) - 2 \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle) \\ & \geq \frac{1}{2\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}) - 2 \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle) \\ & \geq \frac{1}{2\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} 2\tilde{\mu} d(\mathbf{x}, \mathcal{X}^*)^2 = \frac{1}{\gamma^A(\mathbf{x}, \mathbf{x}^*)^2} \tilde{\mu} \|\mathbf{x} - \mathbf{x}^*\|^2. \end{aligned}$$

We can do this for each non-optimal \mathbf{x} . We thus obtain:

$$\tilde{\mu}_f \geq \inf_{\substack{\mathbf{x}, \mathbf{x}^* \in \mathcal{M} \\ \text{s.t. } \langle \mathbf{r}_{\mathbf{x}}, \mathbf{x}^* - \mathbf{x} \rangle > 0}} \tilde{\mu} \left(\frac{\langle \mathbf{r}_{\mathbf{x}}, \mathbf{s}_f(\mathbf{x}) - \mathbf{v}_f(\mathbf{x}) \rangle}{\langle \mathbf{r}_{\mathbf{x}}, \mathbf{x}^* - \mathbf{x} \rangle} \|\mathbf{x}^* - \mathbf{x}\| \right)^2. \quad (40)$$

And we are back to the same situation as in the proof of our earlier Theorem 6, the only change being that we now have equation (25) holding for the general strong convexity constant $\tilde{\mu}$ instead of its classical analogue μ . \square

Having this tool at hand, the linear convergence of all Frank-Wolfe algorithm variants now holds with the earlier μ_f^A complexity constant replaced with $\tilde{\mu}_f$. The factor of 2 in the denominator of (39) is to ensure the same scaling.

Again, as we have shown in Theorem 10, we have that our condition $\tilde{\mu}_f > 0$ leading to linear convergence is slightly weaker than generalized strong convexity in the Hoffman sense (it is implied by it).

Theorem 11. *Suppose that f has smoothness constant C_f^A (C_f for FCFW and MNP), as well as generalized geometric strong convexity constant $\tilde{\mu}_f$ as defined in (39).*

Then the suboptimality error h_t of the iterates of all the four variants of the FW algorithm (AFW, FCFW, MNP and PFW) decreases geometrically at each step that is not a drop step nor a swap step (i.e. when $\gamma_t < \gamma_{\max}$), with the same constants as in Theorem 8, except that μ_f^A is replaced by $\tilde{\mu}_f$.

Proof. The proof closely follows the proof of Theorem 8.

We start from the above generalization (39) of the original geometric strong convexity constant (22), and first replace the inf over \mathbf{x} by considering only the choice $\mathbf{x} := \mathbf{x}^{(t)}$, giving

$$\tilde{\mu}_f \leq \sup_{\substack{\mathbf{x}^* \in \mathcal{X}^* \\ \text{s.t. } \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}^* - \mathbf{x}^{(t)} \rangle < 0}} \frac{1}{2\gamma^A(\mathbf{x}^{(t)}, \mathbf{x}^*)^2} (f(\mathbf{x}^*) - f(\mathbf{x}^{(t)}) - 2 \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}^* - \mathbf{x}^{(t)} \rangle). \quad (41)$$

From here, we will now mirror our earlier derivation for an upper bound on the suboptimality as a function of the gap g_t , as given in (27). As an optimal reference point \mathbf{x}^* in (27), we will choose a $\tilde{\mathbf{x}}^*$ attaining the supremum in (41), given $\mathbf{x}^{(t)}$.

We again employ the ‘step-size quantity’ $\bar{\gamma} := \gamma^A(\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^*)$ as defined in (21). Using (41), we have

$$\begin{aligned} 2\bar{\gamma}^2 \tilde{\mu}_f &\leq f(\mathbf{x}^*) - f(\mathbf{x}^{(t)}) + 2 \left\langle -\nabla f(\mathbf{x}^{(t)}), \tilde{\mathbf{x}}^* - \mathbf{x}^{(t)} \right\rangle \\ &= -h_t + 2\bar{\gamma} \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{s}_f(\mathbf{x}^{(t)}) - \mathbf{v}_f(\mathbf{x}^{(t)}) \right\rangle \\ &\leq -h_t + 2\bar{\gamma} \left\langle -\nabla f(\mathbf{x}^{(t)}), \mathbf{s}_t - \mathbf{v}_t \right\rangle \\ &= -h_t + 2\bar{\gamma} g_t, \end{aligned} \tag{42}$$

Therefore $h_t \leq -\frac{\hat{\gamma}^2}{2} \tilde{\mu}_f + \hat{\gamma} g_t$ when writing $\hat{\gamma} := 2\bar{\gamma}$, which is always upper bounded¹⁶ by

$$h_t \leq \frac{g_t^2}{2\tilde{\mu}_f}. \tag{43}$$

which is exactly the bound (28) as in the classical case, with the denominator being $2\tilde{\mu}_f$ instead of $2\mu_f^A$.

From here, the proof of the main convergence Theorem 8 continues without modification, using $\tilde{\mu}_f$ instead of μ_f^A . \square

Supplementary References

- [37] H. Allende, E. Frandi, R. Nanculef, and C. Sartori. A novel Frank-Wolfe algorithm. Analysis and applications to large-scale SVM training. *arXiv:1304.1014v2*, 2013.
- [38] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, second edition, 1999.
- [39] D. Chakrabarty, P. Jain, and P. Kothari. Provable submodular minimization using Wolfe’s algorithm. In *NIPS*. 2014.
- [40] S. Iwata. A faster scaling algorithm for minimizing submodular functions. In *Integer Programming and Combinatorial Optimization*, volume 2337 of *Lecture Notes in Computer Science*, pages 1–8. 2002.
- [41] G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, 1995.

¹⁶Here we have again used the trivial inequality $0 \leq a^2 - 2ab + b^2$ for the choice of numbers $a := \frac{g_t}{\tilde{\mu}_f}$ and $b := \hat{\gamma}$.