# Affine Invariant Analysis of Frank-Wolfe on Strongly Convex Sets

**Thomas Kerdreux** [*1] **Lewis Liu** [*2 3] **Simon Lacoste Julien** [2 3 4 5] **Damien Scieur** [*4 3]

## Abstract

It is known that the Frank-Wolfe (FW) algorithm, which is affine covariant, enjoys faster convergence rates than $\mathcal{O}(1/K)$ when the constraint set is strongly convex. However, these results rely on norm-dependent assumptions, usually incurring non-affine invariant bounds, in contradiction with FW's affine covariant property. In this work, we introduce new structural assumptions on the problem (such as the directional smoothness) and derive an affine invariant, norm-independent analysis of Frank-Wolfe. We show that our rates are better than any other known convergence rates of FW in this setting. Based on our analysis, we propose an affine invariant backtracking line-search. Interestingly, we show that typical backtracking line-searches using smoothness of the objective function present similar performances than its affine invariant counterpart, despite using affine dependent norms in the step size's computation.

## 1. Introduction

Conditional Gradient algorithms, a.k.a. Frank-Wolfe (FW) algorithms (Frank et al., 1956), form a class of first-order methods solving optimization problems such as

$$\min_{x \in \mathcal{C}} f(x), \quad \mathcal{C} \text{ convex and compact.} \tag{1}$$

FW algorithms decompose non-linear constrained problems into a series of linear problems on the original constraint set, *i.e.* linear minimization oracles (LMO). They form a practical family of algorithms (Jaggi, 2013; Bojanowski et al., 2014; Alayrac et al., 2016; Seguin et al., 2016; Peyre et al., 2017; Miech et al., 2018; Lacoste-Julien et al., 2015; Courty et al., 2016; Paty & Cuturi, 2019; Luise et al., 2019; Combettes & Pokutta, 2021); however, many open questions remain in designing such optimal algorithmic schemes (*e.g.* (Braun et al., 2017; Kerdreux et al., 2018; Braun et al., 2019; Combettes & Pokutta, 2020; Carderera & Pokutta, 2020; Mortagy et al., 2020; Combettes et al., 2020; Bomze et al., 2021)) and in their theoretical understanding.

**Algorithm 1** Frank-Wolfe Algorithm

**Input:** $x_0 \in \mathcal{C}$.
1: **for** $k = 0, 1, \dots, K$ **do**
2: $v_k \in \operatorname{argmax}_{v \in \mathcal{C}} \langle -\nabla f(x_k), v - x_k \rangle$ ▷ LMO
3: $\gamma_k = \operatorname{argmin}_{\gamma \in [0,1]} f(x_k + \gamma(v_k - x_k))$ ▷ Line-search
4: $x_{k+1} = (1 - \gamma_t)x_k + \gamma_k v_k$ ▷ Convex update
5: **end for**

Besides, with the appropriate line-search, the iterates of the FW are *affine covariant* under the affine transformation $y = Bx + b$ of problem (1),

$$\min_{y \in \tilde{\mathcal{C}} = B\mathcal{C} + b} \tilde{f}(y) \overset{\text{def}}{=} f(B^{-1}(y - b)), \quad B \text{ invertible.} \tag{2}$$

**Definition 1.1 (Affine covariance)** *An algorithm is affine covariant when its iterates $(x_k)$ (resp. $(y_k)$) for problem (1) (resp. (2)) satisfy*

$$y_k = Bx_k + b.$$

In other words, the behavior of Algorithm 1 is insensitive to affine transformations or re-parametrization of the space. This means that, ideally, the theoretical rate for an affine covariant algorithm should be *affine invariant*.

The original Frank-Wolfe algorithm (Algorithm 1) generally enjoys a slow sublinear rate $\mathcal{O}(1/K)$ over general compact convex sets and smooth convex functions (Jaggi, 2013). In that setting, (Clarkson, 2010; Jaggi, 2013) define a modulus of smoothness that leads to affine invariant analysis of the Frank-Wolfe algorithm, matching with the affine covariant behavior of the algorithm. Importantly, this analysis is better than any other known *best norm-dependent* analysis. (By *best norm-dependent analysis*, we refer to

[*]Equal contribution [1]Zuse Institute, Berlin [2]Département d'informatique et de recherche opérationnelle (DIRO), Université de Montréal [3]Mila, Montréal [4]Samsung SAIT AI Lab, Montréal [5]Canada CIFAR AI Chair. Correspondence to: Thomas Kerdreux <thomaskerdreux@gmail.com>, Damien Scieur <damien.scieur@gmail.com>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

the norm that minimizes the convergence rate of the theoretical analysis that depend on norms, see, e.g., (Lan, 2013, 3.13.)).

**Definition 1.2 (norm-independence)** *A quantity is norm-independent if it does not depend on the choice of a norm.*

**Counterexample.** *The condition number in optimization – the ratio between the smoothness and the strong convexity constants (Nesterov, 2013) – is norm-dependent. Therefore, algorithms whose rate depends on this condition number may be faster if the choice of the norm makes the condition number closer to* 1.

**Example.** *The curvature constant $C_f$ (Jaggi, 2013) is defined by the ratio*

$$C_f \overset{def}{=} \sup_{\substack{x,\, v \in \mathcal{C} \\ \gamma \in [0,1] \\ y = x + \gamma(v-x)}} \frac{1}{\gamma^2}\Big[f(y) - f(x) - \langle y - x; \nabla f(x)\rangle\Big],$$

*where $C$ is a compact, convex set. Since this ratio does not involve any norm, it is therefore norm-independent.*

Affine invariance and norm-independence are closely related, although they are quite different in nature. We discuss extensively their common points and differences in Appendix A. However, since the FW algorithm is affine invariant and norm-independent, its analysis should ideally also satisfy such properties.

Many works have then sought to find structural assumptions and algorithmic modifications that accelerate this sublinear rate of $\mathcal{O}(1/K)$. The strong convexity of the set (or more generally uniform convexity, see (Kerdreux et al., 2021b;a)) is one of such structural assumptions which lead to various accelerated convergence rates, like linear convergence rates when the unconstrained optimum is outside the constraint set (Levitin & Polyak, 1966; Demyanov & Rubinov, 1970; Dunn, 1979; Rector-Brooks et al., 2019) or sublinear rates $\mathcal{O}(1/K^2)$ when the function is also strongly convex but without restrictions on the position of the optimum (Garber & Hazan, 2015). However, to the best of our knowledge, there exists no norm-independent affine invariant analysis for these accelerated regimes.

In these "non affine invariant" analyses, structural assumptions like the $L$-smoothness (Definition 1.3) of $f$ and the $\alpha$-strong convexity of $\mathcal{C}$ (Definition 1.4) lead to accelerated convergence rate of the Frank-Wolfe algorithm, but are typically conditioned on parameters $L, \alpha$ and others, which depend on a particular choice of a norm. This is surprising given that the Frank-Wolfe algorithm (under appropriate line-search) does not depend on any norm choice.

Recall that the smoothness of a function and the strong convexity of a set are defined as follows.

**Definition 1.3** *The function $f$ is **smooth** over $\mathcal{C}$ if there exists a constant $L > 0$ such that, for any $x, y \in \mathcal{C}$, we have*

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x\rangle + \frac{L}{2}\|x - y\|^2. \tag{3}$$

**Definition 1.4** *A set $\mathcal{C}$ is $\alpha$-**strongly convex** if, for any $(x, y) \in \mathcal{C}$, $\gamma \in [0, 1]$ and $\|z\| \leq 1$, we have*

$$\gamma x + (1-\gamma)y + \frac{\alpha}{2}\gamma(1-\gamma)\|x - y\|^2 z \in \mathcal{C}. \tag{4}$$

Obtaining *practical* accelerated affine invariant rates is hard, as an affine invariant step size is required. Indeed, some adaptive step sizes rely on theoretical affine invariant quantities which are in general not accessible. Therefore, by practical, we consider rates that can be achieved without a deep knowledge of the problem structure and constants.

While the smoothness of a function is quite a standard assumption, the strong convexity of a set is a rather strong assumption. Nevertheless, strong convexity of sets are common in machine learning applications. We can cite, for instance, $\ell_p$ norms (common regularization in machine learning problems or action set in online learning) (Bubeck et al., 2018; Kerdreux et al., 2021c; Wang et al., 2021), matrix Schatten norms (Braverman et al., 2020), and matrix group norms (Kakade et al., 2012).

For instance, scheduled step sizes, *e.g.* $\gamma_k = \frac{2}{k+2}$, makes the Frank-Wolfe algorithm practically affine covariant, yet they do not capture the best accelerated convergence regimes of Frank-Wolfe on strongly convex sets (note, however, the recent proof of an accelerated asymptotic $\mathcal{O}(1/T^2)$ rate of vanilla Frank-Wolfe for specific scheduled step sizes (Bach, 2020)). Exact line-search guarantees a practically affine covariant algorithm while capturing accelerated convergence regimes but significantly increases the time to perform a single iteration. Finally, it is possible to use backtracking line-search such as (Pedregosa et al., 2020). Unfortunately, backtracking techniques rely on the choice of a specific norm, thus breaking affine invariance of the algorithm.

This raises naturally the following questions:

*Can we derive norm-independent, affine invariant rates for Frank-Wolfe on strongly convex sets?*

*Can we design an affine invariant backtracking line-search for Frank-Wolfe algorithms?*

This work provides a positive answer to these questions, by proposing the following contributions.

| Related Work | $\mathcal{C}$ | Str. cvx. $f$ | $x^*$ | Algo | Step size | Rate |
|---|---|---|---|---|---|---|
| Clarkson (2010) | Simplex | ✗ | Any | FW | Scheduled | $\mathcal{O}(1/K)$ |
| Jaggi (2013) | Convex | ✗ | Any | FW | Scheduled | $\mathcal{O}(1/K)$ |
| Lacoste-Julien & Jaggi (2013) | Any | ✓ | Interior | FW | Exact ls | Linear |
| Jaggi & Lacoste-Julien (2015) | Polytope | ✓ | Any | Corr. FW | Exact ls | Linear |
| Gutman & Pena (2020) | | | | | | |
| **Our work** | Strongly cvx | ✗ | $\nabla f(x^\star) \neq 0$ | FW | Backtracking ls | Linear |
| | Strongly cvx | ✓ | Any | FW | Backtracking ls | $\mathcal{O}(1/K^2)$ |

*Table 1.* Existing *affine invariant* analysis of Frank-Wolfe for smooth convex functions under different schemes.

**Strong convexity.** The strong convexity assumption is to be taken in a broad sense. In (Lacoste-Julien & Jaggi, 2013; Jaggi & Lacoste-Julien, 2015), the authors consider "generalized geometric strong convexity" (see their Eq. 39), an affine invariant measure of (generalized) strong convexity, while (Gutman & Pena, 2020) consider strongly convex functions relative to a pair $(\mathcal{C}, \omega)$ where $\omega$ is a distance-like function. In our work, we do not directly assume strong convexity, but the *directional smoothness* of the function (see later Definition 4.1), whose constant is bounded if various assumptions are satisfied for problem (1) (Theorem 4.4).

**Step size.** By *scheduled* step sizes, we consider, for instance, the classical $\gamma_k = \frac{2}{k+2}$. We denote by *exact-line search* when the optimal step size depends on an unknown affine invariant quantity, whose accessible upper-bounds are affine dependent (thus breaking the affine invariance of FW).

**Contributions.** In this paper, **1)** we conduct affine invariant analysis of the Frank-Wolfe Algorithm 1, when the function $f$ is smooth and the set $\mathcal{C}$ is strongly convex. Our affine invariant conditioning is better than any norm-dependent analysis. Additionally, we point out that there is likely a positive gap between our constant and the optimal norm-dependent bound, given that ours are not restricted to a choice of same norms for different parameters in the bound. In specific, we introduce new structural assumptions extending the class of problems for which such accelerated regimes hold in the case of Frank-Wolfe, called *directionally smooth functions w.r.t. a specified direction* $\delta$. Based on this definition, **2)** we propose an affine invariant backtracking line-search for finding the optimal step size, which achieves the best of two worlds in theory and practice. Finally, **3)** we show that existing backtracking line-search methods, which use a specific norm, converges surprisingly to the optimal norm-independent, affine invariant step size. This implies that affine dependent and affine invariant backtracking techniques perform similarly.

**Outline.** In Section 2, we motivate the need for norm-independent affine invariant analysis of Frank-Wolfe on strongly convex sets. In Section 3 and 4, we introduce the structural assumptions on the optimization problem that we will consider for analysing Frank-Wolfe. In Section 5 we detail our affine invariant analysis of Frank-Wolfe on strongly convex set. In Section 6 and 7 we provide a backtracking line-search that directly estimate the affine invariant quantities we developed and we explain how it relates with existing ones. We conclude in Section 8 with numerical experiments.

**Related Work.** Other linear convergence rates of Frank-Wolfe algorithms exists with best affine invariant analysis. For instance, corrective variants of Frank-Wolfe exhibit (affine invariant) linear convergence rates when the constraint set is a polytope (Lacoste-Julien & Jaggi, 2013; Jaggi & Lacoste-Julien, 2015) and the objective function is (generally) strongly convex. See Table 1 for a review of all affine invariant analyses of Frank-Wolfe algorithms.

These affine invariant analyses emphasize that there is no specific choice of norm to be made in Frank-Wolfe algorithms as well as there is no need for affine preconditionners. Frank-Wolfe algorithms are arguably *free-of-choice* methods, *i.e.* little needs to be known on the optimization problem's structures to obtain the accelerated regimes. This is in line with recent works showing that the Frank-Wolfe methods exhibit accelerated adaptive behavior under a variety of structural constraints of (1) which depend on inaccessible parameters (Kerdreux, 2020), *e.g.* Hölderian Error Bounds on $f$ (Kerdreux et al., 2019; Xu & Yang, 2018; Rinaldi & Zeffiro, 2020) or local uniform convexity of $\mathcal{C}$ (Kerdreux et al., 2021b).

Our affine invariant analyses introduce constants seeking to characterize structural properties without a specific choice of norm, even the best (inaccessible) one (see Appendix A for an in-depth discussion). This has been the basis for works extending the accelerated convergence analysis to non-smooth or non-strongly convex functions (Pena, 2019; Gutman & Pena, 2020), which then explore new structural assumptions on $f$.

**Gap between affine invariant and best-norm analysis.** We point out that, in general, affine invariance does not imply optimality. For instance, even if designing norms that produce affine invariant rates is possible (d'Aspremont et al., 2018), this does not imply that such rates will match the result of our norm-independent, affine invariant analy-

sis. In this paper, we show that our affine invariant constant is always better than norm-dependent ones, even after taking the best norm.

Furthermore, it is still an open question if there is a gap between affine invariant rates and the best-norm rate comprising of norm-dependent parameters (such as smoothness, strong convexity and the lower bound of gradient norms). We believe this may be the case, since, the best-norm rate implicitly impose the same norm on all the parameters in the bound, while our affine invariant constant is free of such constraints.

To conclude, we highlight that characterizing the gap between affine invariant and best norm analysis is an interesting and challenging problem in the literature. See Appendix A for more details and examples on this problem.

**Notations.** For a norm $\|\cdot\|$, we write $\|\cdot\|_\star$ its dual norm. Our ambient space is $\mathbb{R}^d$.

## 2. Norm-Dependent Analysis of FW

It is known that when the function is *smooth* (Definition 1.3), the set is *strongly-convex* (Definition 1.4) and the gradient is lower bounded $\|\nabla f(x)\|_\star \geq c > 0$ over the constraint set (i.e., the constraints are active), the Frank-Wolfe algorithm 1 converges linearly (Levitin & Polyak, 1966; Demyanov & Rubinov, 1970; Dunn, 1979), at rate (with $h_k \overset{\text{def}}{=} f(x_k) - f_\star$)

$$h_k \leq \left(\max\left\{\frac{1}{2},\, 1 - \frac{c\alpha}{4L}\right\}\right)^k h_0. \tag{5}$$

see (Garber & Hazan, 2015; Kerdreux et al., 2021b) for more details, and we recall these results in Appendix B, in respectively Lemma B.1 and Corollary B.2. Note that assuming the gradient to be lower bounded means the constraints are tight, i.e., the solution of the unconstrained counterpart lies outside the set of constraints. However, the constants $L$, $\alpha$, and $c$ depend on the choice of the norm for the smoothness and the strong convexity. In contrast, the Frank-Wolfe algorithm and iterates do not depend on such a choice, due to its affine covariance. Therefore, the rate of Algorithm 1 should be affine invariant. Unfortunately, it is possible to show that the known theoretical analyses can be *arbitrarily* bad in the case where the constants $L$, $c$, $\alpha$ depend on "affine variant" norms.

**Example 2.1** *Consider the projection of $\bar{x} : \|\bar{x}\|_2 > 1$,*

$$\min_x f(x) \overset{def}{=} \tfrac{1}{2}\|x - \bar{x}\|_2^2 \quad \textit{such that} \quad \|x\|_2^2 \leq 1.$$

*In such case, we have that $L = 1,\ \alpha = 1$ and $c = \|\bar{x}\|_2 - 1$ ($L$, $\alpha$ and $c$ are defined according to the $\ell_2$ norm, see proof in Appendix B.2). However, if we transform the problem into $\min_y f(By)$, the new constants become*

$$L = \sigma_{\max}(B),\ \ \alpha = \tfrac{\sigma_{\min}(B)}{\sigma_{\max}(B)},\ \ c = \sigma_{\max}(B)(\|\bar{x}\|_2 - 1).$$

*Comparing the rate (5) of the two problems, identical to the eyes of the FW algorithm, we have that*

$$f(x_k) - f^\star \leq \left(1 - \tfrac{\|\bar{x}\|_2 - 1}{4}\right)^k \left(f(x_0) - f^\star\right),$$
$$f(By_k) - f^\star \leq \left(1 - \tfrac{\|\bar{x}\|_2 - 1}{4}\kappa^{-1}(B)\right)^k \left(f(x_0) - f^\star\right),$$

*where $\kappa(B) = \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)}$ is the condition number of $B$. This means we can artificially make a large theoretical upper bound on the rate of convergence by using an ill-conditioned transformation (i.e., $\kappa(B)$ large). However, the speed of convergence of FW iterates are not affected by any linear transformation (dues to their affine covariance), therefore the upper bound will not be representative of the true rate of convergence of FW.*

***Remark.*** *The constants, and therefore the rate, can be improved if we change the norm $\|\cdot\|_2$ into $\|\cdot\|_{2,B^{-1}}$. However, it is usually very hard or impossible to guess what norm will be the best for a specific problem. This is not a problem for FW with exact line-search, as no norm is required. However, in the case of (Garber & Hazan, 2015), the step size (or backtracking line-search) strategy uses $L$, and therefore the rate depends directly on the choice of the norm. Moreover, even if we choose the gauge of the Euclidean ball to measure the function smoothness and the set strong-convexity (becoming, in this case, invariant to affine reparametrization of our problem, see Appendix A)), we do not know how to guarantee it was the optimal choice for this specific problem.*

When the optimum is in the relative interior of any compact set $\mathcal{C}$, FW converges linearly when $f$ is strongly convex (Guélat & Marcotte, 1986; Lacoste-Julien & Jaggi, 2013). On the other hand, linear convergence on strongly convex sets does not require strong convexity of $f$ when the solution of the unconstrained problem lies outside the set (Demyanov & Rubinov, 1970). Our paper hence focuses on extending the analysis where the unconstrained optimum is outside the constraint set (Demyanov & Rubinov, 1970).

These two analysis cover most practical cases, but not the situation where the unconstrained optimum is close to the boundary of $\mathcal{C}$. A recent analysis on strongly convex sets of (Garber & Hazan, 2015) is not restrictive w.r.t. the position of the unconstrained optimum but conservative (convergence rate of $\mathcal{O}(1/K^2)$). It is interesting as it not only deals with the (previously unknown) situation where the unconstrained optimum is on the boundary on $\mathcal{C}$, but also

when it is arbitrarily close to it, leading to poorly conditioned linear convergence regimes. In Appendix E, we provide an affine invariant analysis of (Garber & Hazan, 2015).

## 3. Scaling Inequalities on Strongly Convex Sets

All proofs of Frank-Wolfe methods on strongly convex sets leverage the same property. The *scaling inequality* crucially relates the Frank-Wolfe gap with $\|x_t - v_t\|^2$, see *e.g.* (Kerdreux et al., 2021b, Lemma 2.1.). The scaling inequality is an equivalent characterization of strong convexity of $\mathcal{C}$ (Goncharov & Ivanov, 2017, Theorem 2.1.), but we recall here only the implication that we will need, see (Kerdreux et al., 2021a) for a review of useful properties of uniformly convex sets in machine learning. Importantly, the scaling inequalities motivate the new structural assumptions we present in Section 4 and Appendix E.

**Lemma 3.1 (Norm Scaling Inequality)** *Assume $\mathcal{C}$ is $\alpha$-strongly convex w.r.t. a norm $\|\cdot\|$. Then for any $x \in \mathcal{C}$, $\phi \in \mathbb{R}^d \backslash \{0\}$, and $v_\phi \in \operatorname{argmax}_{v\in\mathcal{C}} \langle \phi, v\rangle$, we have $\phi \in N_{\mathcal{C}}(v_\phi)$ (normal cone) and*

$$\langle \phi,\, v_\phi - x\rangle \geq \tfrac{\alpha}{2}\|\phi\|_\star \|v_\phi - x\|^2. \tag{6}$$

*In particular for any iterate $x_k$ of Frank-Wolfe and its Frank-Wolfe vertex $v_k$ (Line 1 in Algorithm 1), we have*

$$\langle -\nabla f(x_k); v_k - x_k\rangle \geq \tfrac{\alpha}{2}\|\nabla f(x_k)\|_\star \|v_k - x_k\|^2.$$

**Proof.** We start with $v_\phi = \operatorname{argmax}_{v\in\mathcal{C}}\langle \phi;\ v\rangle$. Then, we use the definition of strong convexity of a set,

$$\gamma x + (1-\gamma)v_\phi + \tfrac{\alpha}{2}\gamma(1-\gamma)\|x - v_\phi\|^2 z \in \mathcal{C} \quad \forall z : \|z\| \leq 1.$$

Then, by optimality of $v_\phi$ on $\mathcal{C}$,

$$\langle \phi;\ v_\phi\rangle \geq \langle \phi;\ \gamma x + (1-\gamma)v_\phi + \tfrac{\alpha}{2}\gamma(1-\gamma)\|x - v_\phi\|^2 z\rangle$$

After simplification,

$$\langle \phi;\ v_\phi - x\rangle \geq \tfrac{\alpha}{2}(1-\gamma)\|x - v_\phi\|^2 \langle \phi;\ z\rangle.$$

With $\gamma \to 0$, and after maximizing over $z$, we obtain by definition of $\|\cdot\|_\star$,

$$\langle \phi;\ v_\phi - x\rangle \geq \tfrac{\alpha}{2}\|x - v_\phi\|^2 \|\phi\|_\star,$$

which holds in particular when $\phi = -\nabla f(x)$. ■

These scaling inequalities can take other forms as in the following corollary.

**Corollary 3.2** *Assume $\mathcal{C}$ is $\alpha$-strongly convex w.r.t. $\|\cdot\|$. Consider $(d_1, d_2) \in \mathbb{R}^d$ s.t. $\min\{\|d_1\|_\star, \|d_2\|_\star\} > c > 0$ and let $(x_1, x_2) \in \partial\mathcal{C}$, s.t. $d_i \in N_{\mathcal{C}}(x_i)$ for $i = 1, 2$. Then*

$$\|x_1 - x_2\| \leq \|d_1 - d_2\|_\star / (\alpha c).$$

**Proof.** By applying successively Lemma 3.1, we obtain

$$\langle d_1; x_1 - x_2\rangle \geq \alpha/2 \|d_1\|_\star \|x_1 - x_2\|^2$$
$$\langle d_2; x_2 - x_1\rangle \geq \alpha/2 \|d_2\|_\star \|x_1 - x_2\|^2.$$

We then obtain $\langle d_1 - d_2; x_1 - x_2\rangle \geq \alpha c \|x_1 - x_2\|^2$. Finally, by definition of the dual norm, we conclude that $\alpha c \|x_1 - x_2\| \leq \|d_1 - d_2\|_\star$. ■

This Corollary provides new insights on (Lan, 2013, Algorithm 4). Indeed, it implies that when the set $\mathcal{C}$ is strongly convex and $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_\star > c > 0$, then the strong condition on the Linear Minimization Oracle (Lan, 2013, Equation (4.4.)) is satisfied with $\rho = 1$ and hence PA-CndG (Lan, 2013, Algorithm 4) converges in $\mathcal{O}(1/K^2)$ (Lan, 2013, Corollary 1).

PA-CndG is a Frank-Wolfe type algorithm with the oblivious step-sizes $\frac{2}{k+2}$, hence affine co-variant. Note, however, that the $\mathcal{O}(1/K^2)$ accelerated convergence rate is achieved under the same structural assumption that ensures linear convergence of Frank-Wolfe in (Levitin & Polyak, 1966), which on the other hand, require exact line-search or problem-dependent step-sizes.

## 4. Directional Smoothness

Analyses of Frank-Wolfe algorithm on strongly convex sets show that, when $f$ is convex and smooth, and the unconstrained minima of $f$ are outside of $\mathcal{C}$, there is linear convergence. We hence propose a novel condition that mingles the smoothness of $f$ with the strong convexity of $\mathcal{C}$ when moving in a specific direction $\delta$. We are interested in particular with the FW direction and we will see later that this assumption guarantees a linear convergence rate in this case. We call this condition the *directional smoothness*.

**Definition 4.1** *The function $f$ is directionally smooth with direction function $\delta : \mathcal{C} \to \mathbb{R}^d$ if there exists a constant $\mathcal{L}_{f,\delta} > 0$ s.t. $\forall x \in \mathcal{C}$ and $h > 0$ with $x + h\delta(x) \in \mathcal{C}$,*

$$f\big(x + h\delta(x)\big) \leq f(x) - h\langle -\nabla f(x),\, \delta(x)\rangle + \frac{\mathcal{L}_{f,\delta} h^2}{2}\langle -\nabla f(x),\, \delta(x)\rangle. \tag{7}$$

The rationale of Definition 4.1 is to replace the norm in the usual smoothness condition (Definition 1.3) by a scalar product between the *direction* and the negative gradient, in order to get an affine invariant quantity for the FW direction (see Proposition 4.3 below).

Assuming $\delta(x)$ is a descent direction, i.e., $\langle -\nabla f(x),\, \delta(x)\rangle > 0$, we can obtain a minimization algorithm for $f$, by minimizing (7) over $h$,

$$x_{k+1} = x_k + h_{\text{opt}}\delta(x_k), \quad h_{\text{opt}} = \min\{h_{\max}\ ;\ \mathcal{L}_{f,\delta}^{-1}\}.$$

**Example 4.2** *(Gradient descent on smooth functions) The gradient algorithm uses $\delta(x) = -\nabla f(x)$. In such case, the function is directionally smooth with constant $L$,*

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - h\|\nabla f(x)\|_2^2 + \tfrac{Lh^2}{2}\|\nabla f(x)\|_2^2 \\ &= f(x) + h\left(\tfrac{Lh}{2} - 1\right)\|\nabla f(x)\|_2^2. \end{aligned}$$

*The best $h$ is given by $h_{opt} = \frac{1}{L}$, which is also the optimal one (Nesterov, 2013).*

The advantage of directional smoothness is its affine invariance in the case where $\delta(x)$ is the FW step.

**Proposition 4.3 (Affine Invariance of $\mathcal{L}_{f,\delta}$)** *If $\delta(x)$ is affine covariant (e.g. the FW direction $\delta(x) \overset{def}{=} v(x) - x$), then $\mathcal{L}_{f,\delta}$ in (7) is invariant to an affine transformation of the constraint set (proof in Appendix B.3).*

The next theorem shows that, in the case of the FW algorithm, the directional smoothness constant is bounded if the function is smooth and the set is strongly convex for any norm $\|\cdot\|$.

**Theorem 4.4 (Directional Smoothness of FW)** *Consider the function $f$, smooth w.r.t. the norm $\|\cdot\|$, with constant $L_{\|\cdot\|}$, and the set $\mathcal{C}$, strongly convex with constant $\alpha_{\|\cdot\|}$. Let $\delta(x) = v(x) - x$, $v(x)$ being the FW vertex*

$$v(x) \overset{def}{=} \operatorname*{argmin}_{v\in\mathcal{C}} \langle \nabla f(x),\, v\rangle. \tag{8}$$

*Then, if $\|\nabla f(x)\|_\star > c_{\|\cdot\|}$ for all $x \in \mathcal{C}$ and some $c_{\|\cdot\|} > 0$, the function $f(x)$ is directionally smooth w.r.t. to $\delta$, with*

$$\mathcal{L}_{f,\delta} \leq 2\frac{L_{\|\cdot\|}}{c_{\|\cdot\|}\alpha_{\|\cdot\|}}. \tag{9}$$

**Proof.** See Appendix B.4 for the proof. ■

## 5. Affine Invariant Linear Rates

With the directional smoothness constant $\mathcal{L}_{f.\delta}$ (affine invariant when $\delta$ is the FW direction), Theorem 5.1 shows an affine invariant linear rate of convergence of FW, generalizing existing convergence results of Frank-Wolfe on strongly convex sets (Levitin & Polyak, 1966; Demyanov & Rubinov, 1970; Dunn, 1979).

**Theorem 5.1 (Affine Invariant Linear Rates)** *Assume $f$ is a convex function and directionally smooth with direction function $\delta$ with constant $\mathcal{L}_{f,\delta}$. Then, the FW Algorithm 1 with step size*

$$h_{opt} = \min\left\{1,\ \tfrac{1}{\mathcal{L}_{f,\delta}}\right\}, \quad \text{with } \delta = v(x) - x,$$

*or with line-search, where $v(x)$ is the FW vertex (8), converges linearly, at rate*

$$f(x_k) - f_\star \leq \max\left\{\tfrac{1}{2},\ 1 - \tfrac{1}{2\mathcal{L}_{f,\delta}}\right\}(f(x_{k-1}) - f_\star).$$

**Proof.** We start with the directional smoothness assumption. For $0 < h \leq 1$,

$$f(x_{k+1}) \leq f(x_k) + \left(h - \tfrac{\mathcal{L}_{f,\delta}h^2}{2}\right)\langle \nabla f(x_k),\, \delta(x_k)\rangle$$

After minimization, we have two possibilities: $h_{\text{opt}} = \frac{1}{\mathcal{L}_{f,\delta}}$ or $h_{\text{opt}} = 1$. In the first case, we obtain

$$f(x_{k+1}) \leq f(x_k) + \tfrac{1}{2\mathcal{L}_{f,\delta}}\langle \nabla f(x_k),\, \delta(x_k)\rangle$$

Notice that the scalar product in the right-hand-side is the negative dual gap of Frank-Wolfe, that satisfies

$$\langle \nabla f(x_k),\, v(x) - x\rangle \leq -(f(x_k) - f_\star),$$

which gives the desired result. The second case follows immediately. ■

This provides an affine invariant analysis of the linear convergence regimes of FW on strongly convex sets.

The next corollary shows that the directional constant in Theorem 5.1 is bounded by (9) w.r.t. the norm $\|\cdot\|$ that gives the best ratio.

**Corollary 5.2** *Write $\Omega$ the set of norms in $\mathbb{R}^d$. Then, the rate of convergence using directional smoothness is at least better than the previously known, norm-dependent rate,*

$$1 - \frac{1}{2\mathcal{L}_{f,\delta}} \leq 1 - \frac{1}{4\min_{\|\cdot\|\in\Omega}\frac{L_{\|\cdot\|}}{c_{\|\cdot\|}\alpha_{\|\cdot\|}}},$$

*where $L_{\|\cdot\|}$ is the smoothness constant of the function $f$, $\alpha_{\|\cdot\|}$ the strong convexity of the set $\mathcal{C}$ and $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_\star = c_{\|\cdot\|} > 0$.*

**Proof.** The proof is immediate by noticing that the FW algorithm do not use $\|\cdot\|$, therefore we can choose the best $\|\cdot\|$ in Theorem 4.4. ■

In Appendix E, we provide an affine invariant analysis without restriction on the position of the optimum, *i.e.* the $\mathcal{O}(1/K^2)$ analysis in (Garber & Hazan, 2015). We define (Definition E.1) a similar property to the directional smoothness that additionally accounts for the strong convexity of $f$. We choose to present the affine invariant analysis for the linear convergence in the main body of the paper as it is the one most significant in practice.

## 6. Affine Invariant Backtracking

In previous sections, we proposed new constants to bound the rate of convergence of FW. The significant advantage of these constants is that, like FW, they are independent of any norm. However, the optimal step size of FW needs the knowledge of these constants.

We propose in this section an affine invariant backtracking technique (Algorithm 2), based on directional smoothness. By construction, the technique finds automatically an estimate of the directional smoothness that satisfies

$$\mathcal{L}_k < 2\mathcal{L}_{f,\delta}, \quad k \geq \log_2\left(\frac{\mathcal{L}_0}{\mathcal{L}_{f,\delta}}\right),$$

at the cost of one additional function evaluation per iteration. It is known that such backtracking technique is, in the worst case, two times slower than FW with the optimal, affine invariant stepsize.

**Algorithm 2** Affine invariant backtracking

**Input:** FW vertex $v_k$, point $x_k$, directional smoothness estimate $\mathcal{L}_k$, function $f$.

1: $\mathcal{L} \leftarrow \mathcal{L}_k$. Define the optimal step size and next iterate in the function of the directional Lipchitz constant:

$$\gamma_\star(\mathcal{L}) \stackrel{\text{def}}{=} \min\{\tfrac{1}{\mathcal{L}}, 1\},$$
$$x(\mathcal{L}) \stackrel{\text{def}}{=} (1-\gamma_\star(\mathcal{L}))x_k + \gamma_\star(\mathcal{L})v_k.$$

2: Create the model of $f$ between $x_k$ and $x(\mathcal{L})$ based on equation (7),

$$m(\mathcal{L}) \stackrel{\text{def}}{=} f(x_k) + \gamma_\star(\mathcal{L})\,(1-\gamma_\star(\mathcal{L}))\,\langle \nabla f(x_k),\, v_k - x_k\rangle$$

3: Set the current estimate $\tilde{\mathcal{L}} \stackrel{\text{def}}{=} \frac{\mathcal{L}_k}{2}$.

4: **while** $f(x(\tilde{\mathcal{L}})) > m(\tilde{\mathcal{L}})$ (Sufficient decrease not met because $\tilde{\mathcal{L}}$ is too small) **do**

5: Double the estimate : $\tilde{\mathcal{L}} \leftarrow 2 \cdot \tilde{\mathcal{L}}$.

6: **end while**

**Output:** Estimate $\mathcal{L}_{k+1} = \tilde{\mathcal{L}}$, iterate $x_{k+1} = x(\tilde{\mathcal{L}})$

## 7. Why Backtracking FW with Norms is so Efficient?

The step size strategy in Frank-Wolfe usually drives its practical efficiency. Sometimes, setting the step size optimally w.r.t. the theoretical analysis may be suboptimal in practice. Recently, Pedregosa et al. (2020) analyze the rate of the Frank-Wolfe algorithm for smooth function, using *backtracking line search*, described in Algorithm 3, Appendix D.

Algorithm 3 in Appendix D is adaptive to the local smoothness constant, and ensures $L_{k+1} < 2L_f$, $L_f$ being the smoothness constant of the function in the $\ell_2$ norm. Pedregosa et al. (2020) observed that the estimate of the Lipchitz constant is often significantly smaller than the theoretical one; they wrote: *"We compared the average Lipschitz estimate $L_t$ and the $L$, the gradient's Lipschitz constant. We found that across all datasets the former was more than an order of magnitude smaller, highlighting the need to use a local estimate of the Lipschitz constant to use a large step size."*

With our analysis, however, we can explain why the estimate of the smoothness constant is much better than the theoretical one. The answer is simple:

*Despite using a non-affine invariant bound, the step size resulting from the estimation of the Lipchitz constant via the backtracking line-search is at worst four times smaller than the theoretical affine invariant stepsize.*

**Proposition 7.1** *Let $f$ be directionally smooth, and let $L(x) = \frac{\mathcal{L}_{f,\delta}\langle \nabla f(x), \delta(x)\rangle}{\|\delta(x)\|_2^2}$. Assume $L(x)$ locally approximately constant, i.e., there exists $k_{\min}$, $k_{\max}$ such that, for $L_{loc} = \max_i L(x_i)$,*

$$\frac{L_{loc}}{2} < L(x_k) \leq L_{loc}, \quad k \in [k_{\min}, k_{\max}].$$

*In this case, the norm-dependent backtracking line-search Algorithm 3 finds*

$$L_k < 2L_{loc}, \;\; k = \left\lceil k_{\min} + \log_2 \frac{L_{k_{\min}}}{L_{loc}} \right\rceil, \ldots, k_{\max},$$

*and its step size $(\gamma_\star)_k$ satisfies*

$$\min\left\{1, \frac{1}{4\mathcal{L}_{f,\delta}}\right\} \leq (\gamma_\star)_k.$$

**Proof.** See Appendix B.5 for the full proof.

**Proof sketch.** The constant $L_{\text{loc}}$ can be seen as the local Lipchitz constant. Indeed, if we write the upper bound given by the directional smoothness, we have

$$f(x) + h\langle \nabla f(x),\, \delta(x)\rangle + \frac{h^2}{2}\mathcal{L}_{f,\delta}\langle \nabla f(x_k),\, \delta(x_k)\rangle$$
$$= f(x) + h\langle \nabla f(x),\, \delta(x)\rangle + L(x)\frac{h^2}{2}\|\delta(x)\|_2^2,$$

where the right-hand-side corresponds to the definition of smoothness (3) at $y = x + \delta(x)$ with a variable constant $L(x)$. The parameter $L(x)$ can thus be seen as a "local Lipchitz constant". If $L(x)$ remains approximately constant, the backtracking line-search will eventually finds an estimation $L_k \leq 2L_{\text{loc}}$. Therefore, with the norm-depend backtracking line-search, the step size will be at worst 4 times smaller than the one of the affine invariant fixed-step strategy. ■

Therefore, the optimal step size from the backtracking line-search with the $\ell_2$ norm is *exactly* the optimal affine invariant step size of our affine invariant analysis from Theorem 5.1.

In conclusion, *even if we use non-affine invariant norms* to find the smoothness constant, surprisingly, *the backtracking procedure finds the optimal, affine invariant step size*.

## 8. Illustrative Experiments

**Quadratic / logistic regression.** We consider the constrained quadratic and logistic regression problem,

$$\min_{x\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} l(a_i^T x, y_i), \tag{10}$$

where $l$ is the quadratic or the logistic loss. Here we adopt the $\ell_2$-ball, defined as

$$\mathcal{C} = \{x : \|x\|_2 \leq R\}, \quad R > 0.$$

Specifically, we compare our affine invariant backtracking method in Algorithm 2 against the naive FW Algorithm 1 with step size $1/L$ (Demyanov & Rubinov, 1970) and back-tracking FW (Pedregosa et al., 2020) on the Madelon dataset (Guyon et al., 2007). The results are shown in Figure 2. In detail, we set $R$ such that the unconstrained optimum $\boldsymbol{x}^*$ satisfies $\|\boldsymbol{x}^*\|_2 = 1.1R$, and the initial iterate $\boldsymbol{x}_0 = \mathbf{0}$. As predicted by our theory, the affine invariant algorithm performs well at the beginning, but after a few iterations the two backtracking techniques behave similarly.

**Projection.** We solve here the projection problem described in Example 2.1, for two cases of $B$: One that corresponds to the original problem, i.e. $B = I$, the second one where $B$ is an ill-conditioned matrix (with the condition number $\kappa(B) = 10^6$). The vector $x_0$ is random in the $\ell_2$ ball, and $\bar{x} = \mathbf{1}_d \cdot (1.1/\sqrt{d})$. We report the results in Figure 1. We compare the standard FW algorithm for smooth functions with step size $1/L$, the FW with backtracking line-search (Algorithm 3) and FW with affine invariant backtracking technique (Algorithm 2). If the problem is well-conditioned ($\kappa(B) = 1$), all methods perform similarly. This is not the case, however, for the ill-conditioned setting, where the FW with no adaptive step size converges extremely slowly compared to the two other methods. We also see that the affine invariant backtracking converges quicker than the standard backtracking. This is explained by the fact that the latter takes a longer time to find the right constant $L_k$, while $\mathcal{L}_k$ remains untouched after an affine transformation.

## 9. Conclusion

In this paper, our theoretical convergence results on strongly convex sets complete the series of accelerated affine invariant analyses of Frank-Wolfe algorithms. To obtain these, we formulate a new structural assumption, the directional smoothness, which we will explore more systematically in future works. Also, we present a new affine invariant backtracking line-search method based on directional smoothness. Within our framework of analysis, we provide a new explanation for the reasons behind the efficiency of the existing backtracking line search, and we show theoretically and experimentally they also find affine invariant step sizes.

## Acknowledgments

This research was partially supported by the Canada CIFAR AI Chair Program. Simon Lacoste-Julien is a CIFAR Fellow in the Learning in Machines & Brains program. Research reported in this paper was also partially supported through the Research Campus Modal funded by the German Federal Ministry of Education and Research (fund numbers 05M14ZAM,05M20ZBM) as well as the Deutsche Forschungsgemeinschaft (DFG) through the DFG Cluster of Excellence MATH+.

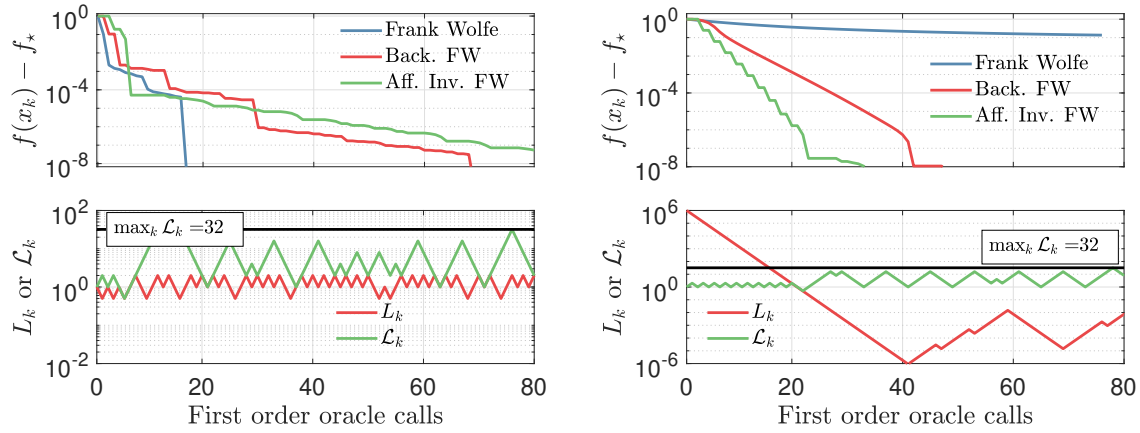


*Figure 1.* Comparison of FW variants on the projection problem. Left: $B = I$, Right: $\kappa(B) = 10^6$. The top row is the gap $f_k - f^*$, and the bottom row corresponds to the estimation of the directional-smoothness constant $\mathcal{L}_k$ or the smoothness constant $L_k$, where the black line report the maximum value of $\mathcal{L}_k$. The reason why adaptive FW methods are slower in the left figure is because, in the worst case, the number of iterations to reach a certain precision can be up to four times larger than the worst-case bound on non-adaptive methods. We clearly see that the directional smoothness parameter $\mathcal{L}_{f,\delta}$ is affine invariant, as its estimate is $\max_k \mathcal{L}_k = 32$ in both scenarios.

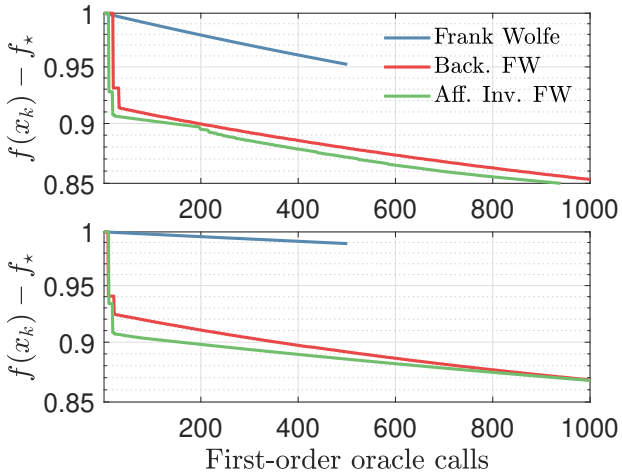


*Figure 2.* Classification problem on Madelon dataset, with (*Top*) Quadratic loss and (*Bottom*) Logistic loss.

## References


Alayrac, J.-B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., and Lacoste-Julien, S. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4575–4583, 2016.

Bach, F. On the effectiveness of Richardson extrapolation in machine learning. *arXiv preprint arXiv:2002.02835*, 2020.

Bauschke, H. H., Bolte, J., and Teboulle, M. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*. Springer, 2014.

Bomze, I. M., Rinaldi, F., and Zeffiro, D. Fast cluster detection in networks by first-order optimization. *arXiv preprint arXiv:2103.15907*, 2021.

Braun, G., Pokutta, S., and Zink, D. Lazifying conditional gradient algorithms. *Proceedings of ICML*, 2017.

Braun, G., Pokutta, S., Tu, D., and Wright, S. Blended conditonal gradients. In *International Conference on Machine Learning*, pp. 735–743. PMLR, 2019.

Braverman, V., Krauthgamer, R., Krishnan, A., and Sinoff, R. Schatten norms in matrix streams: Hello sparsity, goodbye dimension. In *International Conference on Machine Learning*, pp. 1100–1110. PMLR, 2020.

Bubeck, S., Cohen, M., and Li, Y. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pp. 111–127. PMLR, 2018.

Carderera, A. and Pokutta, S. Second-order conditional gradients. *arXiv preprint arXiv:2002.08907*, 2020.

Clarkson, K. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

Combettes, C. and Pokutta, S. Boosting Frank-Wolfe by chasing gradients. In *International Conference on Machine Learning*, pp. 2111–2121. PMLR, 2020.

Combettes, C. W. and Pokutta, S. Complexity of linear minimization and projection on some sets. *arXiv:2101.10040*, 2021.

Combettes, C. W., Spiegel, C., and Pokutta, S. Projection-free adaptive gradients for large-scale optimization. *arXiv:2009.14114*, 2020.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39 (9):1853–1865, 2016.

d'Aspremont, A., Guzman, C., and Jaggi, M. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

Demyanov, V. F. and Rubinov, A. M. Approximate methods in optimization problems. *Modern Analytic and Computational Methods in Science and Mathematics*, 1970.

Dunn, J. C. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Garber, D. and Hazan, E. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *32nd International Conference on Machine Learning, ICML 2015*, 2015.

Goncharov, V. V. and Ivanov, G. E. Strong and weak convexity of closed sets in a Hilbert space. In *Operations research, engineering, and cyber security*, pp. 259–297. Springer, 2017.

Guélat, J. and Marcotte, P. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 1986.

Gutman, D. H. and Pena, J. F. The condition number of a function relative to a set. *Mathematical Programming*, pp. 1–40, 2020.

Guyon, I., Li, J., Mader, T., Pletscher, P. A., Schneider, G., and Uhr, M. Competitive baseline methods set new standards for the nips 2003 feature selection benchmark. *Pattern recognition letters*, 28(12):1438–1444, 2007.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, number CONF, pp. 427–435, 2013.

Jaggi, M. and Lacoste-Julien, S. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28, 2015.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

Kerdreux, T. *Accelerating conditional gradient methods*. PhD thesis, Université Paris sciences et lettres, 2020.

Kerdreux, T., Pedregosa, F., and d'Aspremont, A. Frank-Wolfe with subsampling oracle. In *International Conference on Machine Learning*, pp. 2591–2600. PMLR, 2018.

Kerdreux, T., d'Aspremont, A., and Pokutta, S. Restarting Frank-Wolfe. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

Kerdreux, T., d'Aspremont, A., and Pokutta, S. Local and global uniform convexity conditions. *arXiv preprint arXiv:2102.05134*, 2021a.

Kerdreux, T., d'Aspremont, A., and Pokutta, S. Projection-free optimization on uniformly convex sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 19–27. PMLR, 2021b.

Kerdreux, T., Roux, C., d'Aspremont, A., and Pokutta, S. Linear bandits on uniformly convex sets. *arXiv preprint arXiv:2103.05907*, 2021c.

Lacoste-Julien, S. and Jaggi, M. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

Lacoste-Julien, S., Lindsten, F., and Bach, F. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pp. 544–552. PMLR, 2015.

Lan, G. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Levitin, E. S. and Polyak, B. T. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

Lu, H., Freund, R. M., and Nesterov, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. In *Advances in Neural Information Processing Systems*, pp. 9318–9329, 2019.

Miech, A., Laptev, I., and Sivic, J. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

Molinaro, M. Curvature of feasible sets in offline and online optimization. *arXiv:2002.03213*, 2020.

Mortagy, H., Gupta, S., and Pokutta, S. Walking in the shadow: A new perspective on descent directions for constrained minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Paty, F.-P. and Cuturi, M. Subspace robust wasserstein distances. In *International Conference on Machine Learning*, pp. 5072–5081. PMLR, 2019.

Pedregosa, F., Negiar, G., Askari, A., and Jaggi, M. Linearly convergent Frank-Wolfe with backtracking linesearch. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–10. PMLR, 2020.

Pena, J. Generalized conditional subgradient and generalized mirror descent: duality, convergence, and symmetry. *arXiv preprint arXiv:1903.00459*, 2019.

Peyre, J., Sivic, J., Laptev, I., and Schmid, C. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5179–5188, 2017.

Rector-Brooks, J., Wang, J.-K., and Mozafari, B. Revisiting projection-free optimization for strongly convex constraint sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1576–1583, 2019.

Rinaldi, F. and Zeffiro, D. A unifying framework for the analysis of projection-free first-order methods under a sufficient slope condition. *arXiv:2008.09781*, 2020.

Rockafellar, R. T. *Convex analysis*. Princeton university press, 1970.

Seguin, G., Bojanowski, P., Lajugie, R., and Laptev, I. Instance-level video segmentation from object tracks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Wang, H., Yang, X., and Deng, X. A hybrid first-order method for nonconvex lp-ball constrained optimization. *arXiv preprint arXiv:2104.04400*, 2021.

Xu, Y. and Yang, T. Frank-Wolfe method is automatically adaptive to error bound condition. *arXiv:1810.04765*, 2018.

# A. Affine Invariance of Some Norm-dependent Conditioning in Frank-Wolfe

When solving a smooth constrained convex problem with the vanilla Frank-Wolfe, there are a few known convergence regimes. For instance, the general case of compact convex set with $\mathcal{O}(1/K)$ rates, the case of strongly convex sets with $\mathcal{O}(1/K^2)$ rates when the $f$ is strongly convex and the linear regime when $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_\star > 0$. For the $\mathcal{O}(1/K)$ regime, (Jaggi, 2013) proposes an affine invariant analysis that is better than any known convergence rates. When the set is strongly convex, we also propose such an affine invariant analysis.

This section recalls the relation between these best (w.r.t. to known analysis) affine invariant analyses and the best norm-dependent analyses. To our knowledge, it is a (hard) open question to determine for which problems the best norm conditioning is equal to the affine invariant conditioning of (Jaggi, 2013). We also outline that, in some settings, there are some choices of norms for which the norm-dependent analysis is affine invariant, yet not necessarily better than the best norm analysis.

**Best-affine invariant versus best-norm analyses in FW.** For the general $\mathcal{O}(1/K)$ of FW on general compact convex set, Jaggi (2013) proposes an affine invariant analysis conditioned by the curvature constant of $f$ on $\mathcal{C}$

$$C_f \stackrel{\text{def}}{=} \sup_{\substack{x,s\in\mathcal{C}\\ \gamma\in[0,1]\\ y=x+\gamma(v-x)}} \frac{1}{\gamma^2}\Big[f(y) - f(x) - \langle y-x; \nabla f(x)\rangle\Big]. \tag{11}$$

Previously known norm-dependent $\mathcal{O}(1/K)$ analyses of Frank-Wolfe are conditioned by $L_{\|\cdot\|}D^2_{\|\cdot\|}$, where $L_{\|\cdot\|}$ is the smoothness constant (3) of $f$ measured with respect to a norm $\|\cdot\|$ and $D_{\|\cdot\|}$ the diameter of $\mathcal{C}$ measured with the same norm $\|\cdot\|$. The affine invariant analysis of (Jaggi, 2013) is better than any known other analysis is the sense that (see also (Lan, 2013))

$$C_f \leq \min_{w\in\mathcal{F}} L_{\|\cdot\|}D^2_{\|\cdot\|}, \tag{12}$$

where $\mathcal{F}$ denotes the set of norms (Jaggi, 2013, Lemma 7). The right-hand side of inequality (12) corresponds to the best norm-dependent analysis. In this setting, to the best of our knowledge, *it is an open question* to characterize the problems (1) where (12) is tight. Besides, an immediate advantage of the affine invariant conditioning $C_f$ is that it reduces the problem of finding the best norm into finding the constant $C_f$.

In the setting with strongly convex set, when $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_\star > 0$ for some norm, our affine invariant conditioning is also better than the best norm-dependent analysis, i.e., we have

$$\mathcal{L}_{f,\delta} \leq \min_{w\in\mathcal{F}} \frac{L_w}{c_w\alpha_w}, \tag{13}$$

where $\alpha_w$ is the strong convexity constant of $\mathcal{C}$ measured w.r.t. the norm $w$. However, it is also an open question to understand when the inequality is tight. In an effort to address this question, in Appendix C we extend the Definition 1.3 of smoothness of $f$ and the Definition 1.4 of strong convexity of a set $\mathcal{C}$ when measured w.r.t. more generic function than norms. In particular, inequality (13) remains valid on a larger set than $\mathcal{F}$, but failed to prove that using this new set instead of $\mathcal{F}$ lower the minimum in (12).

**Affine invariant analysis with the problem gauge norm.** In some cases, it is easy to design a norm such that the norm-dependent analysis is affine invariant. However, this does not provide indication that the minimum in (12) or (13) is attained for such norm choice. This was studied in depth in (d'Aspremont et al., 2018) for another family of algorithms.

Assume that $\mathcal{C}$ is a centrally symmetric compact convex set with non-empty interior. This restriction is very mild because it corresponds to most machine learning problems we are aware of. In that setting, the gauge function $\|\cdot\|_\mathcal{C}$ of $\mathcal{C}$ is a norm (Rockafellar, 1970, §15)

$$\|x\|_\mathcal{C} \stackrel{\text{def}}{=} \big\{\mu > 0 \mid x \in \mu\mathcal{C}\big\}. \tag{14}$$

d'Aspremont et al. (2018) show that the smoothness and strong convexity constant of $f$ measured with the gauge of the problem constraint set are invariant to an (invertible) affine re-parametrization of the optimization problem. We recall this property in the following lemma

**Lemma A.1** *(d'Aspremont et al., 2018) Consider $\mathcal{Q}$ a compact centrally symmetric convex set with non-empty interior and an invertible matrix $A$. The function $f$ is $L_{\mathcal{Q}}$-smooth on $\mathcal{Q}$ w.r.t. $\|\cdot\|_{\mathcal{Q}}$, i.e. for any $x, y \in \mathcal{Q}$*

$$f(y) \leq f(x) + \langle \nabla f(x); y - x\rangle + \frac{1}{2} L_{\mathcal{Q}} \|\cdot\|_{\mathcal{Q}}^2,$$

*if and only if $f(A\cdot)$ is $L_{\mathcal{Q}}$-smooth on $A^{-1}\mathcal{Q}$ w.r.t. $\|\cdot\|_{A^{-1}\mathcal{Q}}$. Hence, $L_{\mathcal{Q}}$ if invariant to an affine reparametrization of the optimization problem, i.e. affine invariant.*

We can obtain the same result for the strong convexity of the set when measured with the set gauge function.

**Lemma A.2** *Consider $\mathcal{Q}$ a compact centrally symmetric convex set with non-empty interior and an invertible matrix $A$. The compact convex set $\mathcal{C}$ is $\alpha_{\mathcal{Q}}$-strongly convex w.r.t. $\|\cdot\|_{\mathcal{Q}}$ if and only if $A^{-1}\mathcal{C}$ is $\alpha_{\mathcal{Q}}$-strongly convex w.r.t. $\|\cdot\|_{A^{-1}\mathcal{Q}}$. Hence, $\alpha_{\mathcal{Q}}$ is invariant to an affine reparametrization of the optimization problem, i.e., is affine invariant.*

**Proof.** Assume $\mathcal{C}$ is $\alpha_{\mathcal{Q}}$-strongly convex w.r.t. $\|\cdot\|_{\mathcal{Q}}$. Consider $(x, y, z) \in A^{-1}\mathcal{Q}$. We have $Ax, Ay, Az \in \mathcal{Q}$. For $\gamma \in [0, 1]$, we hence have by definition (because $Az \in \mathcal{Q}$ is equivalent to $\|Az\|_{\mathcal{Q}} \leq 1$)

$$\big\|A(\gamma x + (1-\gamma)y + \alpha_{\mathcal{Q}}\gamma(1-\gamma)\|A(x-y)\|_{\mathcal{Q}}^2 z)\big\|_{\mathcal{Q}} \leq 1$$

Then, by definition of the gauge fct $\|\cdot\|_{\mathcal{Q}}$, note that we have $\|Ax\|_{\mathcal{Q}} = \|x\|_{A^{-1}\mathcal{Q}}$. Hence, last inequality transforms in

$$\big\|\gamma x + (1-\gamma)y + \alpha_{\mathcal{Q}}\gamma(1-\gamma)\|x-y\|_{A^{-1}\mathcal{Q}}^2 z\big\|_{A^{-1}\mathcal{Q}} \leq 1.$$

This means that $A^{-1}\mathcal{Q}$ is $\alpha_{\mathcal{Q}}$-strongly convex w.r.t. $\|\cdot\|_{A^{-1}\mathcal{Q}}$. We do the same reasoning for the other implication. ■

It is then now elementary to prove that the various norm conditioning of Frank-Wolfe analysis are affine invariant.

**Proposition A.3 (Set-norm-dependent Conditioning are affine invariant)** *Assume that $\mathcal{C}$ is a centrally symmetric convex compact set with non-empty interior. Consider a smooth convex function $f$. The following quantity is affine invariant*

$$L_{\|\cdot\|_{\mathcal{C}}} D_{\|\cdot\|_{\mathcal{C}}}^2, \tag{15}$$

*where $L_{\|\cdot\|_{\mathcal{C}}}$ is smoothness constant of $f$ on $\mathcal{C}$ measure w.r.t. $\|\cdot\|_{\mathcal{C}}$ and $D_{\|\cdot\|_{\mathcal{C}}}$ is the diameter. Besides, if $\mathcal{C}$ is strongly convex and $c_{\|\cdot\|_{\mathcal{C}}} \stackrel{def}{=} inf_{x\in\mathcal{C}} \|\nabla f(x)\|_{\mathcal{C}}^\star > 0$, then the following quantity is affine invariant*

$$\frac{L_{\|\cdot\|_{\mathcal{C}}}}{c_{\|\cdot\|_{\mathcal{C}}} \alpha_{\|\cdot\|_{\mathcal{C}}}}, \tag{16}$$

*where $\alpha_w$ is the strong convexity constant of $\mathcal{C}$ when measured w.r.t. $\|\cdot\|_{\mathcal{C}}$.*

**Proof.** From Lemma A.1, the quantity $L_{\|\cdot\|_{\mathcal{C}}}$ is affine invariant. Besides, take $(x, y) \in A^{-1}\mathcal{C}$, we have $\|A(x-y)\|_{\mathcal{C}} = \|x-y\|_{A^{-1}\mathcal{C}}$, hence $D_{\|\cdot\|_{\mathcal{C}}}$ is affine invariant and so is the quantity $L_{\|\cdot\|_{\mathcal{C}}} D_{\|\cdot\|_{\mathcal{C}}}^2$.

Now assume that $\mathcal{C}$ is strongly convex, from Lemma A.2, we also have that $\alpha_{\|\cdot\|_{\mathcal{C}}}$ is affine invariant. Let us verify that $c_w$ is invariant to an affine reparametrization of the problem. We have

$$c_{\|\cdot\|_{\mathcal{C}}} = \inf_{x\in\mathcal{C}} \|\nabla f(x)\|_{\mathcal{C}}^\star = \inf_{y\in A^{-1}\mathcal{C}} \|\nabla f(Ay)\|_{\mathcal{C}}^\star = \inf_{y\in A^{-1}\mathcal{C}} \|AA^{-1}\nabla f(Ay)\|_{\mathcal{C}}^\star.$$

Then, since $\|Ax\|_{\mathcal{C}} = \|x\|_{A^{-1}\mathcal{C}}$ and $\nabla f(A\cdot)(y) = A^{-1}\nabla f(Ay)$, we conclude on the affine invariance of $c_{\|\cdot\|_{\mathcal{C}}}$

$$c_{\|\cdot\|_{\mathcal{C}}} = \inf_{y\in A^{-1}\mathcal{C}} \|\nabla f(A\cdot)(y)\|_{A^{-1}\mathcal{C}}^\star = \inf_{y\in A^{-1}\mathcal{C}} \|\nabla f(A\cdot)(y)\|_{A^{-1}\mathcal{C}}^\star = c_{\|\cdot\|_{A^{-1}\mathcal{C}}}.$$

■

With Proposition A.3, we highlight that it is relatively easy to construct affine invariant bounds for Frank-Wolfe in the existing settings: it suffices to consider the gauge norm when measuring the various problem properties. However, we are not aware of any result establishing an equality like

$$\inf_{w \in \mathcal{F}} \frac{L_w}{c_w \alpha_w} = \frac{L_{\|\cdot\|_\mathcal{C}}}{c_{\|\cdot\|_\mathcal{C}} \alpha_{\|\cdot\|_\mathcal{C}}},$$

where $\mathcal{F}$ is the set of norms on $\mathbb{R}^d$. At a high-level, such a result would be surprising since $\|\cdot\|_\mathcal{C}$ is only aware of the geometry of $\mathcal{C}$, without any consideration w.r.t. the smoothness of $f$ over $\mathcal{C}$ for instance.

## B. Missing Proofs

### B.1. Linear Norm-Dependent Proof of FW on Strongly Convex Sets

For completeness, we recall here the linear convergence regime of Frank-Wolfe when the set is strongly convex and there exists $c > 0$ s.t. $\inf_{x \in \mathcal{C}} \|\nabla f(x)\|_\star > c > 0$ (Levitin & Polyak, 1966; Demyanov & Rubinov, 1970; Dunn, 1979). We now recall the proof from (Garber & Hazan, 2015).

**Lemma B.1** *Consider a $L$-smooth convex function $f$ w.r.t. $\|\cdot\|$ and a compact convex set $\mathcal{C}$ is $\alpha$-strongly convex w.r.t. $\|\cdot\|$. Assume that there exist $c_{\|\cdot\|} > 0$ s.t. $\inf_{x \in \mathcal{C}} \|\nabla f(x)\|_\star > c_{\|\cdot\|}$. Then the iterates $(x_k)$ of Frank-Wolfe (Algorithm 1) are such that*

$$f(x_k) - f^* \leq \left(\max\left\{\frac{1}{2}, 1 - \frac{c\alpha}{8L}\right\}\right)^k (f(x_0) - f^*).$$

**Proof.** Consider $x_k$ the current iterate, $v_k$ the Frank-Wolfe vertex and write $h_k = v_k - x_k$. By definition of the strong convexity of $\mathcal{C}$, we have that $(x_k + v_k)/2 + \alpha/8 \|x_k - v_k\|^2 z \in \mathcal{C}$ for any unit vector $z$. Hence, by optimality of $v_k$, we obtain

$$\langle -\nabla f(x_k); v_k - x_k \rangle \geq \langle -\nabla f(x_k); v_k - x_k \rangle / 2 + \alpha/8 \|x_k - v_k\|^2 \langle -\nabla f(x_k); z \rangle.$$

Then with $\langle -\nabla f(x_k); v_k - x_k \rangle \geq h_k$ and the optimal choice of $z$ this transforms in (it corresponds to (Garber & Hazan, 2015, Equation (5)))

$$\langle -\nabla f(x_k); v_k - x_k \rangle \geq h_k/2 + \alpha/8 \|x_k - v_k\|^2 \|\nabla f(x_k)\|_\star.$$

Then, the $L$-smoothness of $f$ gives (with $x_{k+1} = x_k + \gamma_k(v_k - x_k)$)

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma_k \langle -\nabla f(x_k); v_k - x_k \rangle + L\gamma_k^2/2 \|x_k - v_k\|^2 \\ h_{k+1} &\leq h_k(1 - \gamma_k/2) + \left(L\gamma_k^2 - \alpha c/4\gamma_k\right) \|x_k - v_k\|^2/2. \end{aligned} \tag{17}$$

Then, by the optimality of the choice of $\gamma_k$, and distinguishing according the two cases $L \leq \alpha c/4$ or $L > \alpha c/4$, we obtain the result. ■

Now, we simply note that the ratio in Lemma B.1 can be improved by a factor $1/2$ as in (Kerdreux et al., 2021b).

**Corollary B.2** *With the same conditions as in Lemma B.1 we have*

$$f(x_k) - f^* \leq \left(\max\left\{\frac{1}{2}, 1 - \frac{c\alpha}{4L}\right\}\right)^k (f(x_0) - f^*).$$

**Proof.** Let us start the proof of Lemma B.1 from (17) that we can rewrite (using $\langle -\nabla f(x_k); v_k - x_k \rangle \geq h_k$) as

$$h_{k+1} \leq h_k(1 - \gamma_k/2) - \gamma_k/2 \langle -\nabla f(x_k); v_k - x_k \rangle + L\gamma_k^2/2 \|x_k - v_k\|^2.$$

Then, by using the scaling inequality 6, we obtain

$$h_{k+1} \leq h_k(1 - \gamma_k/2) + \left(L\gamma_k^2 - \alpha c/2\gamma_k\right) \|x_k - v_k\|^2/2,$$

which allows to conclude similarly. ■

### B.2. Proof of Example 2.1

**Erratum.** We restate here Example 2.1, as we spotted an error in the main paper. $\sigma_{\max}(B)$ and $\sigma_{\min}(B)$ denote the largest and smallest singular values of $B$.

**Example B.3** *Consider the projection of $\bar{x} : \|\bar{x}\|_2 > 1$,*

$$\min_x f(x) \overset{def}{=} \tfrac{1}{2}\|x - \bar{x}\|_2^2 \quad \textit{such that} \quad \|x\|_2^2 \leq 1.$$

*In such case, we have that $L = 1,\ \alpha = 1$ and $c = \|\bar{x}\|_2 - 1$ ($L$, $\alpha$ and $c$ are defined according to the $\ell_2$ norm, see proof in Appendix B.2). However, if we transform the problem into $\min_{y \in B^{-1}\ell_2(1)} f(By)$, the new constants become*

$$L = \sigma_{\max}^2(B),\ \ \alpha = \tfrac{\sigma_{\min}^2(B)}{\sigma_{\max}(B)},\ \ c = \sigma_{\min}(B)(\|\bar{x}\|_2 - 1).$$

*Comparing the rate (5) of the two problems, identical to the eyes of the FW algorithm, we have that*

$$f(x_k) - f^\star \leq \left(1 - \tfrac{\|\bar{x}\|_2 - 1}{4}\right)^k \left(f(x_0) - f^\star\right),$$
$$f(By_k) - f^\star \leq \left(1 - \tfrac{\|\bar{x}\|_2 - 1}{4}\kappa^{-3}(B)\right)^k \left(f(x_0) - f^\star\right),$$

*where $\kappa(B) = \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)}$ is the condition number of $B$. This means we can artificially make a large theoretical upper bound on the rate of convergence by using an ill-conditioned transformation (i.e., $\kappa(B)$ large). However, the speed of convergence of FW iterates are not affected by any linear transformation (dues to their affine covariance), therefore the upper bound will not be representative of the true rate of convergence of FW.*

Before proving the quantities in the example, we recall an important result from (Garber & Hazan, 2015).

**Lemma B.4** *(Garber & Hazan, 2015, Lemma 2) Let the set $\mathcal{C} = \{x : f(x) \leq 1\}$, where $f$ is $L$-smooth and $\mu$-strongly convex. Then, the set $\mathcal{C}$ is strongly convex with constant $\frac{\mu}{\sqrt{2L}}$.*

**Lemma B.5 (Example 2.1)** *Let $f(x) \overset{def}{=} \frac{1}{2}\|x - \bar{x}\|_2^2$, $B$ an invertible matrix, and $\mathcal{C}$ the Euclidean ball. We write $\sigma_{max}(B)$ the largest singular value of $B$. We have the following properties on $f$, $f(B\cdot)$, $\mathcal{C}$, and $B^{-1}\mathcal{C}$.*

*(a) $f$ is 1-smooth w.r.t. $\|\cdot\|_2$,*

*(b) $f(B\cdot)$ is $\sigma_{max}^2(B)$-smooth w.r.t. $\|\cdot\|_2$,*

*(c) $\mathcal{C} = \{x : \|x\|^2 \leq 1\}$ is 1-strongly convex w.r.t. $\|\cdot\|_2$,*

*(d) $B^{-1}\mathcal{C} = \{y : \|By\|^2 \leq 1\}$ is $\frac{\sigma_{min}^2(B)}{\sigma_{max}(B)}$-strongly convex w.r.t. $\|\cdot\|_2$,*

*(e) $inf_{x \in \mathcal{C}} \|\nabla f(x)\|_2 = \|\bar{x}\|_2 - 1$,*

*(f) $inf_{y \in B^{-1}\mathcal{C}} \|\nabla f(B\cdot)(y)\|_2 = \|B^T\bar{x}\| \left(1 - \frac{1}{\|\bar{x}\|}\right) \geq \sigma_{\min}(B)(\|\bar{x}\| - 1)$.*

**Proof.** [Proof of Example 2.1]

(a) By construction, the function $f(x) = \frac{1}{2}\|x - \bar{x}\|_2^2$ is smooth with constant $L = 1$.

(b) We have $\nabla^2 \frac{1}{2}\|By - \bar{x}\|^2 = B^T B \preceq \sigma_{\max}^2 I$, therefore the function is smooth with constant $\sigma_{\max}^2$.

(c) We have that $\|\cdot\|^2$ is smooth with constant $L = 2$ and strongly convex with constant $\mu = 2$. Therefore, from Lemma B.4, the set is strongly convex with constant $\alpha = 1$.

(d) We have that $\|B\cdot\|^2$ is smooth with constant $L = 2\sigma_{\max}^2(B)$ and strongly convex with constant $\mu = 2\sigma_{\min}^2(B)$. Therefore, from Lemma B.4, the set is strongly convex with constant $\alpha = \frac{\sigma_{\min}^2(B)}{\sigma_{\max}(B)}$.

(e) The solution of $\min_{\|x\|\leq 1} \nabla f(x) = \|x - \bar{x}\|_2$ is simply the $\ell_2$ projection of the point $\bar{x}$ onto the unit euclidean ball, i.e., the solution is achieved at $x^* = \frac{\bar{x}}{\|\bar{x}\|}$. Therefore, the minimum is $\|x^* - \bar{x}\|_2 = \|\bar{x}\| - 1$.

(f) We use the previous solution, and use the fact that $By = x$. Therefore, the solution is achieved at $y^* = B^{-1}x^* = B^{-1}\frac{\bar{x}}{\|\bar{x}\|}$. This gives

$$\begin{aligned}\|\nabla f(By^*)\| &= \|B^T(By^* - \bar{x})\|,\\ &= \|B^T\left(\frac{\bar{x}}{\|\bar{x}\|} - \bar{x}\right)\|,\\ &= \|B^T\bar{x}\|\left(1 - \frac{1}{\bar{x}}\right),\\ &\geq \sigma_{\min}(B)(\|\bar{x}\| - 1).\end{aligned}$$

■

### B.3. Proof of Proposition 4.3

**Proposition B.6 (Affine Invariance of $\mathcal{L}_{f,\delta}$)** *If $\delta(x)$ is affine covariant (e.g. the FW direction $\delta(x) \stackrel{def}{=} v(x) - x$), then $\mathcal{L}_{f,\delta}$ in (7) is invariant to an affine transformation of the constraint set (proof in Appendix B.3).*

**Proof.** We start with the definition of directional smoothness, but with $x \to By$. The upper bound reads

$$f(By) + \left(h - \frac{\mathcal{L}_{f,\delta}h^2}{2}\right)\langle \nabla f(By),\, \delta(By)\rangle$$

Since we assumed $\delta(By)$ affine covariant,

$$\delta(By) = B\tilde{\delta}_{\tilde{\mathcal{C}}}(y).$$

Therefore,

$$f(By) + \left(h - \frac{\mathcal{L}_{f,\delta}h^2}{2}\right)\langle B^T\nabla f(By),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle$$

Since $\nabla \tilde{f}(y) = B^T\nabla f(By)$, we have

$$\tilde{f}(\tilde{y} + h\tilde{\delta}_{\tilde{\mathcal{C}}}(y)) \leq \tilde{f}(y) + \left(h - \frac{\mathcal{L}_{f,\delta}h^2}{2}\right)\langle \nabla \tilde{f}(y),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle$$

This means the function $\tilde{f}$ is directionally smooth with constant $\mathcal{L}_{f,\delta}$, which proves the statement. ■

### B.4. Proof of Theorem 4.4

**Theorem B.7 (Directional Smoothness of FW)** *Consider the function $f$, smooth w.r.t. the norm $\|\cdot\|$, with constant $L_{\|\cdot\|}$, and the set $\mathcal{C}$, strongly convex with constant $\alpha_{\|\cdot\|}$.*
*Let $\delta(x) = v(x) - x$, $v(x)$ being the FW vertex*

$$v(x) \stackrel{def}{=} \operatorname*{argmin}_{v\in\mathcal{C}} \langle \nabla f(x),\, v\rangle. \tag{18}$$

*Then, if $\|\nabla f(x)\|_\star > c_{\|\cdot\|}$ for all $x \in \mathcal{C}$ and some $c_{\|\cdot\|} > 0$, the function $f(x)$ is directionally smooth w.r.t. to $\delta$, with*

$$\mathcal{L}_{f,\delta} \leq 2\frac{L_{\|\cdot\|}}{c_{\|\cdot\|}\alpha_{\|\cdot\|}}. \tag{19}$$

**Proof.** We start by the definition of smooth functions between $x$ and $h\delta(x)$ w.r.t. the norm $\|\cdot\|$. We have for all $0 \leq h \leq 1$

$$f(x + h\delta(x)) \leq f(x) + h\langle \nabla f(x),\, \delta(x)\rangle + \frac{h^2 L_{\|\cdot\|}}{2}\|\delta(x)\|^2.$$

Using the scaling inequality in (6),

$$\langle -\nabla f(x),\, \delta(x)\rangle \geq \frac{\alpha_{\|\cdot\|}}{2}\|\nabla f(x)\|_\star \|\delta(x)\|^2.$$

We hence obtain

$$f(x + h\delta(x)) \leq f(x) + h\langle \nabla f(x),\, \delta(x)\rangle - h^2 L_{\|\cdot\|}\frac{\langle \nabla f(x),\, \delta(x)\rangle}{\alpha_{\|\cdot\|}\|\nabla f(x)\|_\star}.$$

Since $\|\nabla f(x)\|_\star > c_{\|\cdot\|}$ for all $x \in \mathcal{C}$,

$$f(x + h\delta(x)) \leq f(x) + h\langle \nabla f(x),\, \delta(x)\rangle - \frac{h^2}{2}\frac{2L_{\|\cdot\|}}{\alpha_{\|\cdot\|}c_{\|\cdot\|}}\langle \nabla f(x),\, \delta(x)\rangle.$$

which is the definition of directional smoothness. ■

### B.5. Proof of Proposition 7.1

**Proposition B.8** *Let $f$ be directionally smooth, and let $L(x) = \frac{\mathcal{L}_{f,\delta}\langle \nabla f(x),\, \delta(x)\rangle}{\|\delta(x)\|_2^2}$. Assume $L(x)$ locally approximately constant, i.e., there exists $k_{\min}$, $k_{\max}$ such that, for $L_{loc} = \max_i L(x_i)$,*

$$\frac{L_{loc}}{2} < L(x_k) \leq L_{loc}, \quad k \in [k_{\min},\, k_{\max}].$$

*In this case, the norm-dependent backtracking line-search Algorithm 3 finds*

$$L_k < 2L_{loc},\ k = \left\lceil k_{\min} + \log_2 \frac{L_{k_{\min}}}{L_{loc}} \right\rceil, \ldots, k_{\max},$$

*and its step size $(\gamma_\star)_k$ satisfies*

$$\min\left\{1, \frac{1}{4\mathcal{L}_{f,\delta}}\right\} \leq (\gamma_\star)_k.$$

**Proof.** First, we show that, when $f$ is directionally smooth, we have that $f$ is smooth w.r.t. a certain norm $\|\cdot\|$, in the direction $\delta$, with a constant $L(x)$. Indeed,

$$\begin{aligned}
f(x + h\delta(x)) &\leq f(x) + h\nabla f(x)\delta(x) + \frac{h^2}{2}\mathcal{L}_{f,\delta}\langle \nabla f(x_k),\, \delta(x_k)\rangle && \text{(Directional smoothness)}\\
&= f(x) + h\nabla f(x)\delta(x) + \frac{\mathcal{L}_{f,\delta}\langle \nabla f(x_k),\, \delta(x_k)\rangle}{\|\delta(x)\|_2^2}\frac{h^2}{2}\|\delta(x)\|_2^2 &&\\
&= f(x) + h\nabla f(x)\delta(x) + L(x)\frac{h^2}{2}\|\delta(x)\|_2^2. && \text{(Smoothness with } y = x + \delta(x).)
\end{aligned}$$

Now, consider we start Algorithm 3 with the value $L_{k_{\min}}$, which is the current estimate of the Lipchitz constant using the norm-dependent backtracking technique. To prove that it takes $\log_2 \frac{L_{\min}}{L_{\text{loc}}}$ iterations from $k_{\min}$ to achieve $L_k \leq L_{\text{loc}}$, it suffices to notice that the backtracking line-search will divide $L_{k_{\min}}$ by 2 at each step. Indeed, from the definition of $m$ in Algorithm 3, for $k = k_{\min} \ldots k_{\min} + \log_2 \frac{L_{k_{\min}}}{L_{\text{loc}}}$,

$$m_k(L_k) = f(x + h\delta(x)) \leq f(x) + h\nabla f(x)\delta(x) + L_k\frac{h^2}{2}\|\delta(x)\|_2^2 \geq f(x) + h\nabla f(x)\delta(x) + L(x)\tfrac{h^2}{2}\|\delta(x)\|_2^2.$$

since $L_k = \frac{L_{k_{\min}}}{2^i} \geq L_{loc} \geq L(x)$.

We now have $L_{k+i} \leq 2L_{\text{loc}}$ for $i \geq \log_2 \frac{L_{k_{\min}}}{L_{\text{loc}}}$, and $L(x) \leq L_{k+i}$, otherwise the condition in Step 4 in Algorithm 3 is not met. Moreover, we have $\frac{L_{loc}}{2} < L(x)$ by assumption. All together,

$$\frac{L_{loc}}{2} < L(x_k) \leq L_k \leq 2L_{\text{loc}}.$$

Now, if we take the expression of the stepsize $\gamma_*(L_k)$ from Algorithm 3, we have

$$\begin{aligned}
\gamma_\star(L) &\stackrel{\text{def}}{=} \min\left\{\langle\frac{-\nabla f(x_k),\, \delta_k\rangle}{L_k\|\delta_k\|^2}, 1\right\}, \\
&\geq \min\left\{\frac{\langle-\nabla f(x_k),\, \delta_k\rangle}{2L_{\text{loc}}\|\delta_k\|^2}, 1\right\}, \\
&\geq \min\left\{\frac{\langle-\nabla f(x_k),\, \delta_k\rangle}{4L(x_k)\|\delta_k\|^2}, 1\right\}.
\end{aligned}$$

We obtain the result by replacing $L(x)$ with its expression. ■

## C. Strong Convexity of Sets with Asymmetric Distance Functions

A limitation in the definition of smoothness of a function (Definition 1.3) and the strong convexity of a set (Definition 1.4) is the presence of the norm in their definition, whose constants may be dependent on affine transformation of the space (see Example 2.1). Technically, the notion of norm in the definition of smoothness and strong convexity of a function can be extended to the concept of distance-generating function, for instance using Bregman divergence (Bauschke et al., 2017; Lu et al., 2018) or gauge functions (d'Aspremont et al., 2018).

Although is it classical to use different distance-generating functions $\omega$ (that satisfies Assumption C.1 below) in place of a norm $\|\cdot\|$ to characterize the smoothness of a function, we are not aware of such analysis for strongly convex sets. We believe that such analysis may exist, but for completeness we propose here an extension of the strong convexity of a set w.r.t. a distance function $\omega$.

**Assumption C.1** *The function $\omega(\cdot)$ satisfies*

- $\omega(x) = 0 \quad \Leftrightarrow \quad x = 0$,
- ***Positivity:*** $\omega(x) \geq 0$,
- ***Triangular Inequality:*** $\omega(x + y) \leq \omega(x) + \omega(y)$
- ***Positive homogeneity:*** $\omega(\gamma x) = \gamma\omega(x)$, $\gamma \geq 0$,
- ***Bounded asymmetry:*** $\max_x \frac{\omega(x)}{\omega(-x)} \leq \kappa_\omega$.

Since $\omega(x)$ is convex by the triangle inequality, we define the dual distance

$$\omega_*(v) = \max_{x:\omega(x)\leq 1} \langle v, x\rangle. \tag{20}$$

**Remark C.2** *Usually, extensions of smoothness of a function use Bregman divergences (see e.g. (Lu et al., 2018; Bauschke et al., 2017)). However, the assumption that the distance-generating function is positively homogeneous is crucial in our analysis, which is unfortunately, not satisfied for most Bregman divergences.*

For instance, gauge functions as defined in (14) satisfy Assumption C.1 if the set $\mathcal{Q}$ is convex and compact, and contains $0$ in its interior. Usually, most works using gauge function assume that the set $\mathcal{Q}$ is *centrally symmetric* (d'Aspremont et al., 2018; Molinaro, 2020), which add the assumption that

$$\omega(x) = \omega(-x).$$

In that case, the gauge function is a norm (Rockafellar, 1970, Theorem 15.2.). This restriction reasonably covers most of the practical settings we encounter in machine learning. However, removing symmetry extends non-trivially the definition of strongly convex sets w.r.t. the distance function $\omega$. We will now introduce the necessary concepts to show that Theorem 4.4 is also valid when considering the set of all $\omega$ satisfying Assumptions C.1. This result hints at a possible gap between the norm-dependent affine invariant analysis we provide and the best norm analysis of Frank-Wolfe.

We now recall the definitions of smoothness and strong convexity of a function w.r.t. a distance function $\omega$.

**Definition C.3** *A function $f$ is smooth (resp. strongly convex) w.r.t. the distance function $\omega$ if, for a constant $L_\omega > 0$ (resp. $\mu_\omega > 0$), the function satisfies*

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x\rangle + \frac{L_\omega}{2}\omega^2(y-x), \tag{21}$$

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x\rangle + \frac{\mu_\omega}{2}\omega^2(y-x). \tag{22}$$

**Definition C.4** *A set $\mathcal{C}$ is $\alpha_\omega$-strongly convex w.r.t. $\omega$ if, for any $x, y \in \mathcal{C}$ and $\gamma \in [0,1]$, we have*

$$z_\gamma + \alpha_\omega \gamma(1-\gamma)\frac{(1-\gamma)\omega^2(x-y) + \gamma\omega^2(y-x)}{2} z \in \mathcal{C},$$

*where $z_\gamma = \gamma x + (1-\gamma)y$, for all $z$ such that $\omega(z) \leq 1$.*

This definition extends the one of strongly convex sets with a general distance function that may not be a norm, see Definition 1.4. We can also extend the scaling inequality (6) to strongly convex sets with generic distance functions.

**Lemma C.5 (Distance Scaling Inequality)** *Assume $\mathcal{C}$ is $\alpha_\omega$-strongly convex w.r.t. $\omega$. Then for any $x \in \mathcal{C}$, $\phi \in \mathbb{R}^d \backslash \{0\}$, and $v_\phi \in \operatorname{argmax}_{v\in\mathcal{C}}\langle \phi, v\rangle$, we have $\phi \in N_{\mathcal{C}}(v_\phi)$ (normal cone) and*

$$\langle \phi,\, v_\phi - x\rangle \geq \frac{\alpha_\omega}{2}\omega_*\big(\phi\big)\omega^2(v_\phi - x). \tag{23}$$

*In particular for any iterate $x_k$ of Frank-Wolfe and its Frank-Wolfe vertex $v_k$ (Line 1 in Algorithm 1), we have*

$$\langle -\nabla f(x_k); v_k - x_k\rangle \geq \frac{\alpha_\omega}{2}\omega_*\big(-\nabla f(x_k)\big)\omega^2(v_k - x_k).$$

**Proof.** We start with $v_\phi = \operatorname{argmax}_{v\in\mathcal{C}}\langle \phi;\ v\rangle$. Then, we use the definition of strong convexity of a set,

$$\gamma x + (1-\gamma)v_\phi + \alpha_\omega\gamma(1-\gamma)D_\gamma(x - v_\phi)z \in \mathcal{C} \quad \forall z : \omega(z) \leq 1.$$

where $D_\gamma(x-y) \overset{\text{def}}{=} \frac{\gamma\omega^2(x-y)+(1-\gamma)\omega^2(y-x)}{2}$. Then, by optimality of $v_\phi$,

$$\langle \phi;\ v_\phi\rangle \geq \langle \phi;\ \gamma x + (1-\gamma)v_\phi + \alpha_\omega\gamma(1-\gamma)D_\gamma(x-v_\phi)z\rangle$$

After simplification,

$$\langle \phi;\ v_\phi - x\rangle \geq \alpha_\omega(1-\gamma)D_\gamma(x-v_\phi)\langle \phi;\ z\rangle,$$

which holds in particular when $\phi = -\nabla f(x)$, $\gamma = 0$ and $z$ being the argmax (see (20)). ■

Let us now introduce the following results that extends known properties from smooth and strongly convex sets.

**Proposition C.6** *If $f$ is strongly convex w.r.t. the distance function $\omega$, then for $\gamma \in [0,1]$ we have*

$$f(\gamma x + (1-\gamma)y) + \mu\gamma(1-\gamma)\frac{\gamma\omega^2(x-y) + (1-\gamma)\omega^2(y-x)}{2} \leq \gamma f(x) + (1-\gamma)f(y)$$

**Proof.** Let $z_\gamma = \gamma x + (1-\gamma)y$. We start with the definition,

$$f(z_\gamma) + \langle \nabla f(z_\gamma),\, x - z_\gamma\rangle + \frac{\mu}{2}\omega^2(x - z_\gamma) \le f(x)$$
$$f(z_\gamma) + \langle \nabla f(z_\gamma),\, y - z_\gamma\rangle + \frac{\mu}{2}\omega^2(y - z_\gamma) \le f(y)$$

After multiplying by $\gamma$ and $1-\gamma$ and adding the two inequalities, we have

$$f(z_\gamma) + \mu\frac{\gamma\omega^2(x - z_\gamma) + (1-\gamma)\omega^2(y - z_\gamma)}{2} \le \gamma f(x) + (1-\gamma)f(y)$$

Since $\omega^2(x - z_\gamma) = (1-\gamma)^2\omega^2(y - x)$, and $\omega^2(y - z_\gamma) = \gamma^2\omega^2(x - y)$, we obtain the desired result. ■

**Proposition C.7** *If $f$ is convex and smooth w.r.t. the distance function $\omega$, then it holds that*

$$\frac{1}{2L}\omega_*^2\big(\nabla f(x) - \nabla f(y)\big) \le f(y) - f(x) - \langle \nabla f(x),\, y - x\rangle$$

*where $\omega_*$ is the dual of the function $\omega$, written*

$$\omega_*(v) \stackrel{def}{=} \max_{s:\omega(s)\le 1} \langle v,\, s\rangle.$$

In particular, Proposition C.7 implies that, if $f$ has a minimum $x_\star$, then

$$\frac{1}{2L}\omega_*^2\big(-\nabla f(y)\big) \le f(y) - f(x_\star) \tag{24}$$

**Proof.** Let the function $\phi(y) = f(y) - \langle \nabla f(x), y\rangle$. This function is, by construction, smooth. Moreover, $\min_y \phi(y)$ is attained when $y = x$. Since the function is smooth,

$$\min_y \phi(y) \le \min_y \phi(z) + \langle \nabla\phi(z),\, y - z\rangle + \frac{L}{2}\omega^2(y - z)$$

Let $\beta u = y - z$, where $\omega(u) = 1$ and $\beta \ge 0$. Then,

$$\min_y \phi(y) \le \min_{\beta,u} \phi(z) + \beta\langle \nabla\phi(z),\, u\rangle + \frac{\beta^2 L}{2}$$

The minimum can be split into two minimization problems,

$$\min_y \phi(y) \le \phi(z) + \min_{\beta\ge 0}\left(\frac{\beta^2 L}{2} - \beta \max_{u:\omega(u)\le 1}\langle -\nabla\phi(z),\, u\rangle\right).$$

By definition of the dual of $\omega$,

$$\min_y \phi(y) \le \phi(z) + \min_{\beta\ge 0}\left(\frac{\beta^2 L}{2} - \beta\omega_*\big(-\nabla\phi(z)\big)\right).$$

Now, we can solve over $\beta$, which gives us

$$\min_y \phi(y) \le \phi(z) - \frac{1}{2L}\omega_*^2\big(-\nabla\phi(z)\big).$$

Replacing the minimum by $\phi(x)$, and $\phi$ by its expression, we get

$$f(x) - \langle \nabla f(x),\, x\rangle \le f(z) - \langle \nabla f(x),\, z\rangle - \frac{1}{2L}\omega_*^2\big(\nabla f(x) - \nabla f(z)\big).$$

After reorganization, we get the desired result. ■

With Definition C.4, the level sets of smooth and strongly convex functions are also strongly convex sets when the function $\omega$ is used. This result corresponds *exactly* to the one of (Journée et al., 2010, Theorem 12), when we use $\omega = \|\cdot\|_2$.

**Lemma C.8 (Strong Convexity of Sets)** *Let $f$ be a $L$-smooth and $\mu$-strongly convex function w.r.t. $\omega$. Then, the set*

$$\mathcal{C} = \{x : f(x) - f_\star \leq R\}$$

*is $\alpha$-strongly convex w.r.t. $\omega$, with $\alpha = \frac{\mu_\omega}{\kappa_\omega \sqrt{2L_\omega R}}$.*

**Proof.**

Consider the set

$$\mathcal{C} = \{x : f(x) - f_\star \leq R\}$$

Let $x,\ y \in \mathcal{C}$. Let $z_\gamma = \gamma x + (1-\gamma)y$, and consider the point $z_\gamma + u$. We have that

$$\begin{aligned}
f(z_\gamma + u) - f_\star &\leq f(z_\gamma) - f_\star + \langle \nabla f(z_\gamma),\, u\rangle + \frac{L}{2}\omega^2(u), \\
&\leq f(z_\gamma) - f_\star + \omega(-u) \max_{v:\omega(v)\leq 1} \langle -\nabla f(z_\gamma),\, v\rangle + \frac{L}{2}\omega^2(u), \\
&= f(z_\gamma) - f_\star + \omega(-u)\omega_*\big(-\nabla f(z_\gamma)\big) + \frac{L}{2}\omega^2(u), \\
&\leq f(z_\gamma) - f_\star + \kappa_\omega \omega(u)\sqrt{2L(f(z_\gamma) - f_\star)} + \frac{L}{2}\omega^2(u).
\end{aligned}$$

Therefore, to satisfy $f(z_\gamma + u) - f_\star \leq R$, we need to ensure that

$$\underbrace{f(z_\gamma) - f_\star - R}_{=\omega} + \underbrace{\kappa_\omega \sqrt{2L(f(z_\gamma) - f_\star)}}_{=\beta}\omega(u) + \frac{L}{2}\omega^2(u) \leq 0$$

Solving the problem in $\omega(u)$ gives

$$\omega(u) \leq \frac{-\beta + \sqrt{\beta^2 - 2L\omega}}{L}$$

We have that

$$\beta^2 - 2L\omega = 2L\left((f(z_\gamma) - f_\star)(\kappa_\omega^2 - 1) + R\right)$$

Therefore,

$$\omega(u) \leq \sqrt{2}\frac{-\kappa_\omega\sqrt{(f(z_\gamma) - f_\star)} + \sqrt{(f(z_\gamma) - f_\star)(\kappa_\omega^2 - 1) + R}}{\sqrt{L}}$$

However, since the function is strongly convex,

$$f(z_\gamma) - f_\star \leq \underbrace{\gamma f(x) + (1-\gamma)f(y) - f_\star}_{\leq R} - \mu\gamma(1-\gamma)\frac{\gamma\omega^2(x-y) + (1-\gamma)\omega^2(y-x)}{2}$$

Let $D_\gamma = \gamma(1-\gamma)\frac{\gamma\omega^2(x-y) + (1-\gamma)\omega^2(y-x)}{2}$. The inequality now reads

$$f(z_\gamma) - f_\star \leq R - \mu D_\gamma. \tag{25}$$

Therefore, the condition on $\omega$ becomes

$$\omega(u) \leq \sqrt{2}\frac{-\kappa_\omega\sqrt{R - \mu D_\gamma} + \sqrt{(R - \mu D_\gamma)(\kappa_\omega^2 - 1) + R}}{\sqrt{L}}$$

which gives

$$\omega(u) \leq \frac{\kappa_\omega\sqrt{2}}{\sqrt{L}}\left(-\sqrt{R - \mu D_\gamma} + \sqrt{R - \left(1 - \frac{1}{\kappa_\omega^2}\right)\mu D_\gamma}\right) \tag{26}$$

To simplify the expression in parenthesis, we multiply and divide by the conjugate of the square roots to get:

$$\left(-\sqrt{R-\mu D_\gamma}+\sqrt{R-\left(1-\frac{1}{\kappa_\omega^2}\right)\mu D_\gamma}\right)=\frac{R-\left(1-\frac{1}{\kappa_\omega^2}\right)\mu D_\gamma-(R-\mu D_\gamma)}{\sqrt{R-\mu D_\gamma}+\sqrt{R-\left(1-\frac{1}{\kappa_\omega^2}\right)\mu D_\gamma}}$$
$$\geq\frac{1}{\kappa_\omega^2 2\sqrt{R}}.$$

We can thus strengthen the condition (26) to:

$$\omega(u)\leq\frac{\mu D_\gamma}{\kappa_\omega\sqrt{2LR}}.$$

As the definition of a strongly convex set requires $\omega(u)\leq\alpha_\omega D_\gamma$, we conclude that the level set is strongly convex with at least the constant $\alpha_\omega=\frac{\mu}{\kappa_\omega\sqrt{2LR}}$. ■

We now provide the analog of Theorem 4.4 when the assumptions of $f$ and $\mathcal{C}$ are with respect to a distance generating function $\omega$ instead of a norm.

**Theorem C.9** *Consider the function $f$, smooth w.r.t. the distance function $\omega$, with constant $L_\omega$, and the set $\mathcal{C}$, strongly convex with constant $\alpha_\omega$. Let $\delta(x)=v(x)-x$, $v(x)$ being the FW vertex*

$$v(x)\stackrel{def}{=}\operatorname*{argmin}_{v\in\mathcal{C}}\langle\nabla f(x),\,v\rangle.$$

*Then, if $\omega_*(-\nabla f(x))>c_\omega$ for all $x\in\mathcal{C}$, the function $f(x)$ is directionally smooth w.r.t. to $\omega$, with constant*

$$\mathcal{L}_{f,\delta}\leq\frac{L_\omega}{c_\omega\alpha_\omega}. \tag{27}$$

**Proof.** We start by the definition of smooth functions between $x$ and $h\delta(x)$ for the distance function $\omega$. We have for all $0\leq h\leq 1$

$$f(x+h\delta(x))\leq f(x)+h\langle\nabla f(x),\,\delta(x)\rangle+\frac{h^2L_\omega}{2}\omega^2(\delta(x))$$

Using the scaling inequality in (23),

$$\langle-\nabla f(x),\,\delta(x)\rangle\geq\alpha_\omega\omega_*\big(-\nabla f(x)\big)\omega(\delta(x))^2.$$

We hence obtain

$$f(x+h\delta(x))\leq f(x)+h\langle\nabla f(x),\,\delta(x)\rangle-\frac{h^2L_\omega}{2}\frac{\langle\nabla f(x),\,\delta(x)\rangle}{\alpha_\omega\omega_*\big(-\nabla f(x)\big)}.$$

Since $\omega_*(-\nabla f(x))>c_\omega$ for all $x\in\mathcal{C}$,

$$f(x+h\delta(x))\leq f(x)+h\langle\nabla f(x),\,\delta(x)\rangle-\frac{h^2}{2}\frac{L_\omega}{\alpha_\omega c_\omega}\langle\nabla f(x),\,\delta(x)\rangle.$$

which is the definition of directional smoothness. ■

## D. Backtracking Line Search for Frank-Wolfe Steps

**Algorithm 3** Backtracking line-search for smooth functions (Pedregosa et al., 2020)

**Input:** FW vertex $v_k$, point $x_k$, smoothness estimate $L_k$, function $f$.

1: Create the optimal step size and next iterate in the function of the Lipchitz estimate

$$\gamma_\star(L) \stackrel{\text{def}}{=} \min\left\{\frac{-\nabla f(x_k)(v_k - x_k)}{L\|v_k - x_k\|^2}, 1\right\}.$$
$$x(L) \stackrel{\text{def}}{=} (1 - \gamma_\star(L)) + \gamma_\star(L)v_k$$

2: Quadratic model of $f$ between $x_k$ and $x(L)$,

$$m(L) \stackrel{\text{def}}{=} f(x_k) + \langle \nabla f(x_k),\, x(L) - x_k\rangle + \frac{L}{2}\|x(L) - x_k\|^2$$

3: Set the current estimate $\tilde{L} \stackrel{\text{def}}{=} \frac{L_k}{2}$.

4: **while** $f(x(\tilde{L})) > m(\tilde{L})$ (Sufficient decrease not met because $\tilde{L}$ is too small) **do**

5: Double the estimate : $\tilde{L} \leftarrow 2 \cdot \tilde{L}$.

6: **end while**

**Output:** Estimate $L_{k+1} = \tilde{L}$, iterate $x_{k+1} = x(\tilde{L})$

## E. Affine Invariant Analysis without Restriction on Optimum Location

In this section, we propose a modification of the directional smoothness defined in Section 4. This new assumption is the basis to obtain an affine invariant analysis of Frank-Wolfe on a strongly convex set without restriction on the position of the unconstrained optimum of $f$, as recently proposed in Garber & Hazan (2015).

**Outline.** In Theorem E.2, we prove a $\mathcal{O}(1/K^2)$ sublinear convergence rate as in (Garber & Hazan, 2015) when the function is *modified directionally smooth* (Definition E.1). In Theorem E.4, we prove that when $\mathcal{C}$ is strongly convex, and $f$ is smooth and strongly convex, then $f$ is *modified directionally smooth* for the Frank-Wolfe direction with an affine invariant constant leading to better conditioned convergence rates than in (Garber & Hazan, 2015). Finally, in Proposition E.5, we show that the constant of modified directional smoothness is affine invariant.

We now define a modification of directional smoothness. It is a structural assumption on $f$ constrained on $\mathcal{C}$ designed at gathering the strong convexity of $\mathcal{C}$, the smoothness, and the strong convexity of $f$ into a single quantity.

**Definition E.1 (Modified Directional Smoothness)** *Let $x_0 \in \mathcal{C}$. The function $f$ is called modified directionally smooth with direction function $\delta : \mathcal{C} \to \mathbb{R}^N$ if there exists a constant $\tilde{\mathcal{L}}_{f,\delta}(x_0) > 0$ such that $\forall x \in \mathcal{C}$,*

$$f\big(x + h\delta(x)\big) \leq f(x) + h\langle \nabla f(x),\, \delta(x)\rangle - \frac{\tilde{\mathcal{L}}_{f,\delta}(x_0)h^2}{2}\langle \nabla f(x),\, \delta(x)\rangle \sqrt{\frac{f(x_0) - f^*}{f(x) - f^*}}, \tag{28}$$

*for $0 < h < 1$.*

Note that the dependence of $x_0$ in the definition of the modified directional smoothness is an artifact to obtain a dimensionless constant $\tilde{\mathcal{L}}_{f,\delta}(x_0)$.

As in Section 5, the modified directional smoothness constant $\tilde{\mathcal{L}}_{f.\delta}$ is affine invariant in the case where $\delta$ is the FW direction. We now derive an affine invariant accelerated sublinear rate of convergence of Frank-Wolfe providing an affine invariant analysis of (Garber & Hazan, 2015).

**Theorem E.2 (Affine Invariant Accelerated Sublinear Rates)** *Let $x_0 \in \mathcal{C}$ and assume $f$ is a convex function and modified directionally smooth with direction function $\delta$ and constant $\tilde{\mathcal{L}}_{f,\delta}(x_0)$. Then, the iterates $x_k$ for the Frank-Wolfe Algorithm 1 with step size*

$$h_{opt} = \min\left\{1,\ \frac{1}{\tilde{\mathcal{L}}_{f,\delta}(x_0)}\sqrt{\frac{f(x_k) - f^*}{f(x_0) - f^*}}\right\}, \quad \text{with } \delta = v(x) - x,$$

*or with exact line-search, where $v(x)$ is the Frank-Wolfe vertex*

$$v(x) = \operatorname*{argmin}_{v \in \mathcal{C}} \langle \nabla f(x),\, v \rangle,$$

*satisfy*

$$f(x_k) - f^* \leq \frac{4(f(x_0) - f^*) \max\{1,\, 18\tilde{\mathcal{L}}_{f,\delta}^2(x_0)\}}{(k+2)^2} \quad \textit{for } k \geq 0.$$

**Proof.** The proof is similar to that of Theorem 5.1. We hence start with the modified directional smoothness assumption on $f$. For $0 < h < 1$,

$$f\big(x_{k+1}\big) \leq f(x_k) + \left( h - \frac{\tilde{\mathcal{L}}_{f,\delta} h^2}{2} \sqrt{\frac{f(x_0) - f^*}{f(x_k) - f^*}} \right) \langle \nabla f(x_k),\, \delta(x_k) \rangle \tag{29}$$

After minimizing over $h$, we have two possibilities. The case with exact line-search follows immediately after these two cases. In the following, we use the notation $h_k \overset{\text{def}}{=} f(x_k) - f^*$ for the primal suboptimality at $x_k$, and $g_k \overset{\text{def}}{=} \langle -\nabla f(x_k),\, \delta(x_k) \rangle$ for the Frank-Wolfe gap at $x_k$ (and note that $g_k \geq h_k$ by convexity).

**Case 1:** $h_{\text{opt}} = \frac{1}{\tilde{\mathcal{L}}_{f,\delta}(x_0)} \sqrt{\frac{f(x_k) - f^*}{f(x_0) - f^*}}$. In such case, we obtain (subtract $f^*$ on both sides of the inequality)

$$h_{k+1} \leq h_k - \frac{1}{2\tilde{\mathcal{L}}_{f,\delta}} \sqrt{\frac{h_k}{h_0}} g_k,$$

and since the Frank-Wolfe gap $g_k$ upper bounds the primal suboptimality, we obtain

$$h_{k+1} \leq h_k \Big[ 1 - \frac{1}{2\tilde{\mathcal{L}}_{f,\delta} \sqrt{h_0}} \sqrt{h_k} \Big].$$

**Case 2:** With $h_{\text{opt}} = 1$, we have

$$h_{k+1} \leq h_k + \left( 1 - \frac{\mathcal{L}_{f,\delta}}{2} \sqrt{\frac{h_0}{h_k}} \right) g_k.$$

In that case, we have that $\frac{1}{\tilde{\mathcal{L}}_{f,\delta}(x_0)} \sqrt{\frac{h_k}{h_0}} \geq 1$. Hence we obtain

$$h_{k+1} \leq h_k - \tfrac{1}{2} g_k \leq \tfrac{1}{2} h_k$$

Finally, we have the following recursive relation on the sequence of primal suboptimality $(h_k)$:

$$\begin{aligned} h_{k+1} &\leq h_k \cdot \max\left\{ \frac{1}{2},\, 1 - \frac{1}{2\tilde{\mathcal{L}}_{f,\delta} \sqrt{h_0}} \sqrt{h_k} \right\} \\ &= h_k \cdot \max\left\{ \frac{1}{2},\, 1 - M \sqrt{h_k} \right\}, \end{aligned} \tag{30}$$

with $M \overset{\text{def}}{=} \frac{1}{2\tilde{\mathcal{L}}_{f,\delta}(x_0)\sqrt{h_0}}$. The inequality (30) is exactly the same recurrence that was analyzed by Garber & Hazan (2015) (see their Equation (7), with the same notation for $M$), where they have shown a $\mathcal{O}(1/K^2)$ convergence rate. The exact constant is obtained by following the very same proof as (Garber & Hazan, 2015), *i.e.* proving by induction that there exists $C$ such that $h_k \leq C/(k+2)^2$. The base case $k = 0$ can be trivially obtained by letting $C \geq 4h_0$.[1] Their induction step was shown by requiring that $C \geq \frac{18}{M^2}$. Thus using $C = \max\{4h_0, \frac{18}{M^2}\}$ (and re-arranging) proves the statement of our theorem. ■

The following lemma will be used in the proof of the bound on the modified directional smoothness.

[1] Note that Garber & Hazan (2015) use a different argument for the base case, bounding instead $h_1$ with $L \cdot \operatorname{diam}(\mathcal{C})^2/2$, using the Lipschitz smoothness of $f$ (and this would become $C_f/2$ in its affine invariant formulation with $C_f$ as defined by Jaggi (2013)). However, we believe that $h_0$ is usually smaller than $C_f$ in applications, and in any case $h_0$ appears from $1/M^2$ for us, so using our different base case argument is more meaningful.

**Lemma E.3** *Consider a compact convex set $\mathcal{C}$. Assume $f$ is a $\mu_\omega$-strongly convex function with respect to $\omega$. Let $x^*$ be the minimum of $f$ on $\mathcal{C}$. Then, for any $x \in \mathcal{C}$, we have*

$$\omega_*(\nabla f(x)) \geq \sqrt{\frac{\mu_\omega}{2}}\sqrt{f(x) - f(x^*)}. \tag{31}$$

**Proof.** Let $x \in \mathcal{C}$. From Definition C.3, we have that

$$f(x) \geq f(x^*) + \langle \nabla f(x^*),\, x - x^* \rangle + \frac{\mu_\omega}{2}\omega^2(x - x^\star).$$

Hence with the optimality conditions, *i.e.* $\langle \nabla f(x^*),\, x - x^* \rangle \geq 0$, we have

$$f(x) - f(x^*) \geq \frac{\mu_\omega}{2}\omega^2(x - x^*). \tag{32}$$

By convexity of $f$, we have $\langle x - x^*,\, \nabla f(x) \rangle \geq f(x) - f(x^*)$, and by definition of the Fenchel conjugate, we have

$$\omega(x - x^*) \cdot \omega_*(\nabla f(x)) \geq \langle x - x^*,\, \nabla f(x) \rangle \geq f(x) - f(x^*).$$

Hence by plugging (32), we obtain (31). ■

We now prove Theorem E.4 that is similar to Theorem 4.4. It states that in the case of the FW algorithm, the modified directional smoothness constant is bounded if the function is smooth, strongly convex and the set is strongly convex for any distance function $\omega$. It also provides an explicit upper bound on the modified directional smoothness constant. This bound implies that the convergence rate in Theorem E.2 is better conditioned than existing results (Garber & Hazan, 2015).

**Theorem E.4 (Bounds on modified directional smoothness)** *Consider $x_0 \in \mathcal{C}$ and a function $f$, smooth w.r.t. the distance function $\omega$, with constant $L_\omega$, strongly convex w.r.t. the distance function $\omega$, with constant $\mu_\omega$, and the set $\mathcal{C}$, strongly convex with constant $\alpha_\omega$. Let $\delta(x) = v(x) - x$, $v(x)$ being the FW vertex. Then, the function $f(x)$ is modified directionally smooth w.r.t. to $\delta$, with constant*

$$\tilde{\mathcal{L}}_{f,\delta}(x_0) \leq \frac{\kappa_\omega\sqrt{2}L_\omega}{\alpha_\omega\sqrt{\mu_\omega}}\frac{1}{\sqrt{f(x_0) - f^*}}. \tag{33}$$

**Proof.** Let $h \in [0, 1]$. With the smoothness of $f$, we have

$$f(x + h\delta(x)) \leq f(x) - h\langle -\nabla f(x),\, \delta(x) \rangle + \frac{h^2 L_\omega}{2}\omega\big(\delta(x)\big)^2.$$

Recall that when $\delta(x)$ is the Frank-Wolfe direction, we have that the Frank-Wolfe gap $g(x)$ is equal to $\langle -\nabla f(x),\, \delta(x) \rangle$. Also, the scaling inequality for strongly convex sets (Lemma C.5) implies that $\omega(\delta(x))^2 \leq g(x)/(\alpha_\omega\omega^\star(-\nabla f(x)))$, so that

$$f(x + h\delta(x)) \leq f(x) - h\langle -\nabla f(x),\, \delta(x) \rangle + \frac{h^2 L_\omega}{2\alpha_\omega}\frac{g(x)}{\omega^\star(-\nabla f(x))}.$$

Now, it is easy to see from the definition of the dual distance $\omega_*$ that is has the same bounded asymmetry constant as for $\omega$, and thus $\omega^\star(-\nabla f(x)) \geq \frac{1}{\kappa_\omega}\omega^\star(\nabla f(x))$. Thus we apply (31) to obtain:

$$f(x + h\delta(x)) \leq f(x) - hg(x) + \frac{h^2}{2}\frac{\kappa_w\sqrt{2}L_\omega}{\alpha_\omega\sqrt{\mu_\omega}\sqrt{f(x_0) - f^*}}\frac{\sqrt{f(x_0) - f^*}}{\sqrt{f(x) - f^*}}g(x),$$

which implies equation (33). ■

Theorem E.4 shows that the conditioning of convergence with the directional smoothness, which does not depend on any norm choice, in Theorem E.2 is better than conditioning of other analysis (Garber & Hazan, 2015). We now prove that the optimal constant of modified directional smoothness $\tilde{L}_{f,\delta}$ is affine invariant, a result similar to Proposition 4.3 for the directional smoothness constant.

**Proposition E.5 (Affine Invariance of Modified Directional Smoothness)** *Consider $\mathcal{C}$ a compact convex set and $f$ a convex function on $\mathcal{C}$ that is modified directionally smooth w.r.t. $\delta(x)$ with constant $\tilde{\mathcal{L}}_{f,\delta}(x_0)$ (with $x_0 \in \mathcal{C}$). If for any $x \in \mathcal{C}$, $\delta(x)$ is affine covariant (e.g. the Frank-Wolfe direction $\delta(x) \overset{def}{=} v(x) - x$), then the constant $\tilde{\mathcal{L}}_{f,\delta}$ in (28) is affine invariant. In other words, for an invertible matrix $B$, let*

$$\tilde{f}(\cdot) \overset{def}{=} f(B\cdot), \quad \tilde{\delta}_{\tilde{\mathcal{C}}}(\cdot) \overset{def}{=} \delta_{B^{-1}\cdot\mathcal{C}}(\cdot),$$

*then $\tilde{\mathcal{L}}_{\tilde{f},\tilde{\delta}_{\tilde{\mathcal{C}}}}(x_0) = \tilde{\mathcal{L}}_{f,\delta}(y_0)$, where $y_0 \overset{def}{=} B^{-1}x_0$.*

**Proof.** Let $y \in B^{-1}\cdot\mathcal{C}$. Applying the definition of directional smoothness for $f$ at $By$, we obtain

$$f\big(By + h\delta(By)\big) \leq f(By) + h\langle \nabla f(By),\, \delta(By)\rangle - \frac{\tilde{\mathcal{L}}_{f,\delta}(x_0)h^2}{2}\langle \nabla f(By),\, \delta(By)\rangle \sqrt{\frac{f(x_0) - f^*}{f(By) - f^*}}. \tag{34}$$

Similarly to Proposition 4.3, we have that $\nabla \tilde{f}(y) = B^T \nabla f(By)$ and $\delta(By) = B\tilde{\delta}_{\tilde{\mathcal{C}}}(y)$ so that

$$\langle \nabla f(By),\, \delta(By)\rangle = \langle \nabla f(By),\, B\tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle = \langle B^T\nabla f(By),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle = \langle \nabla \tilde{f}(y),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle.$$

Hence (34) and $\tilde{f}^* = f^*$, implies that for any $y \in B^{-1}\cdot\mathcal{C}$

$$\tilde{f}(y + h\tilde{\delta}_{\tilde{\mathcal{C}}}) \leq \tilde{f}(y) + h\langle \nabla \tilde{f}(y),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle - \frac{\tilde{\mathcal{L}}_{f,\delta}(x_0)h^2}{2}\langle \nabla \tilde{f}(y),\, \tilde{\delta}_{\tilde{\mathcal{C}}}(y)\rangle \sqrt{\frac{\tilde{f}(y_0) - \tilde{f}^*}{\tilde{f}(y) - \tilde{f}^*}}.$$

Hence, $\tilde{f}$ is modified directionally smooth on $\tilde{\mathcal{C}} \overset{\text{def}}{=} B^{-1}\cdot\mathcal{C}$ with respect to $\tilde{\delta}_{\tilde{\mathcal{C}}}$ and $\tilde{L}_{\tilde{f},\tilde{\delta}_{\tilde{\mathcal{C}}}}(y_0) \leq \tilde{\mathcal{L}}_{f,\delta}(x_0)$. A similar reasoning concludes that the two constants are equal. ■

## F. Related Work Details

(Lacoste-Julien & Jaggi, 2013) propose an affine invariant analysis of the vanilla Frank-Wolfe algorithm when the unconstrained optimum $x^*$ is in the relative interior of the constraint set $\mathcal{C}$ and $f$ is strongly convex. Hence, the analysis applies when the constraint set is a strongly convex set, and the quantity might be defined in our context. However, the affine invariant constant $\mu_f^{(FW)}$ standing for the strong convexity of $f$ is zero whenever the optimum is not in the relative interior of the constraint set $\mathcal{C}$. Indeed, Equation (3) from (Lacoste-Julien & Jaggi, 2013) define the following affine invariant quantity

$$\mu_f^{(FW)} \overset{\text{def}}{=} \inf_{\substack{x\in\mathcal{C}\setminus\{x^*\},\gamma\in]0,1]\\ \bar{s}=\bar{s}(x,x^*,\mathcal{C})\\ y=x+\gamma(\bar{s}-x)}} \frac{2}{\gamma^2}\big[f(y) - f(x) - \langle \nabla f(x),\, y - x\rangle\big],$$

where $\bar{s}(x, x^*, \mathcal{C}) = \text{ray}(x, x^*) \cap \partial\mathcal{C}$. When $x^* \notin \mathcal{C}$, we have $\mu_f^{(FW)} \leq 0$ since there are some point $x \in \partial\mathcal{C}$ such that $x \in \bar{s}(x, x^*, \mathcal{C})$, and thus we can take $\bar{s} = x$ in the $\inf$, yielding $y = x$ with $\gamma > 0$. This means that the above quantity cannot be easily generalized to the setting we studied in Theorem 4.4 where the unconstrained optimum is assumed to be *outside* of $\mathcal{C}$.