# Projection-Free Optimization on Uniformly Convex Sets

**Thomas Kerdreux**
Technische Universität
Zuse Institute
Berlin, Germany

**Alexandre d'Aspremont**
Ecole Normale Supérieure
CNRS
Paris, France

**Sebastian Pokutta**
Technische Universität
Zuse Institute
Berlin, Germany

## Abstract

The Frank-Wolfe method solves smooth constrained convex optimization problems at a generic sublinear rate of $\mathcal{O}(1/T)$, and it (or its variants) enjoys accelerated convergence rates for two fundamental classes of constraints: polytopes and strongly-convex sets. Uniformly convex sets non-trivially subsume strongly convex sets and form a large variety of *curved* convex sets commonly encountered in machine learning and signal processing. For instance, the $\ell_p$-balls are uniformly convex for all $p > 1$, but strongly convex for $p \in ]1, 2]$ only. We show that these sets systematically induce accelerated convergence rates for the original Frank-Wolfe algorithm, which continuously interpolate between known rates. Our accelerated convergence rates emphasize that it is the curvature of the constraint sets – not just their strong convexity – that leads to accelerated convergence rates. These results also importantly highlight that the Frank-Wolfe algorithm is adaptive to much more generic constraint set structures, thus explaining faster empirical convergence. Finally, we also show accelerated convergence rates when the set is only locally uniformly convex around the optima and provide similar results in online linear optimization.

## 1 Introduction

The Frank-Wolfe method [Frank and Wolfe, 1956] (Algorithm 1) is a projection-free algorithm designed to solve

$$\operatorname*{argmin}_{x \in \mathcal{C}} f(x), \tag{OPT}$$

where $\mathcal{C}$ is a compact convex set and $f$ a smooth convex function. Many recent algorithmic developments in this family of methods are motivated by appealing properties already contained in the original Frank-Wolfe algorithm. Each iteration requires to solve a Linear Minimization Oracle (see line 2 in Algorithm 1), instead of a projection or proximal operation that is not computationally competitive in various settings [Combettes and Pokutta, 2021]. Also, the Frank-Wolfe iterates are convex combinations of extreme points of $\mathcal{C}$, the solutions of the Linear Minimization Oracle. Hence, depending on the extremal structure of $\mathcal{C}$, early iterates may have a specific structure, being, *e.g.*, sparse or low rank for instance, that could be traded-off with the iterate approximation quality of problem (OPT). These fundamental properties are among the main features that contribute to the recent revival and extensions of the Frank-Wolfe algorithm [Clarkson, 2010, Jaggi, 2011] used for instance in large-scale structured prediction [Bojanowski et al., 2014, 2015, Alayrac et al., 2016, Seguin et al., 2016, Miech et al., 2017, Peyre et al., 2017], quadrature rules in RKHS [Bach et al., 2012, Lacoste-Julien et al., 2015, Futami et al., 2019], optimal transport [Courty et al., 2016, Paty and Cuturi, 2019, Luise et al., 2019], and many others.

**Algorithm 1** Frank-Wolfe Algorithm

**Input:** $x_0 \in \mathcal{C}$, $L$ upper bound on the Lipschitz constant.

1: **for** $t = 0, 1, \ldots, T$ **do**
2: $\quad v_t \in \operatorname*{argmax}_{v \in \mathcal{C}} \langle -\nabla f(x_t), v - x_t \rangle$ ▷ LMO
3: $\quad \gamma_t = \operatorname*{argmin}_{\gamma \in [0,1]} \gamma \langle v_t - x_t, \nabla f(x_t) \rangle + \frac{\gamma^2}{2} L \|v_t - x_t\|^2$
4: $\quad x_{t+1} = (1 - \gamma_t) x_t + \gamma_t v_t$
5: **end for**

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

**Uniform Convexity.** Uniform convexity is a global quantification of the curvature of a convex set $\mathcal{C}$. There exists several definitions, see for instance, [Goncharov and Ivanov, 2017, Theorem 2.1.] and [Abernethy et al., 2018, Molinaro, 2020] for the strongly convex case. Here, we focus on the generalization of a classic definition of the strong convexity of a set [Garber and Hazan, 2015].

**Definition 1.1** ($\gamma$ uniform convexity of $\mathcal{C}$)**.** *A closed set $\mathcal{C} \subset \mathbb{R}^d$ is $\gamma_{\mathcal{C}}$-uniformly convex with respect to a norm $\|\cdot\|$, if for any $x, y \in \mathcal{C}$, any $\eta \in [0,1]$ and any $z \in \mathbb{R}^d$ with $\|z\| = 1$, we have*

$$\eta x + (1-\eta)y + \eta(1-\eta)\gamma_{\mathcal{C}}(\|x-y\|)z \in \mathcal{C},$$

*where $\gamma_{\mathcal{C}}(\cdot) \geq 0$ is a non-decreasing function. In particular when there exists $\alpha > 0$ and $q > 0$ such that $\gamma_{\mathcal{C}}(r) \geq \alpha r^q$, we say that $\mathcal{C}$ is $(\alpha, q)$-uniformly convex or $q$-uniformly convex.*

*A set is $\alpha$-strongly convex if and only if it is $(\alpha, 2)$-uniformly convex.*

The uniform convexity assumption strengthens the convexity property of $\mathcal{C}$ that any line segment between two points is included in $\mathcal{C}$. It requires a scaled unit ball to fit in $\mathcal{C}$ and results in curved sets. Two common families of uniformly convex sets are the $\ell_p$-balls and $p$-Schatten balls which are uniformly convex for any $p > 1$ but strongly convex for $p \in ]1, 2]$ only, *i.e.* 2-uniformly convex sets for $p \in ]1, 2]$.

**Convergence Rates for Frank-Wolfe.** The Frank-Wolfe algorithm admits a tight [Canon and Cullum, 1968, Jaggi, 2013, Lan, 2013] general sublinear convergence rate of $\mathcal{O}(1/T)$ when $\mathcal{C}$ is a compact convex set and $f$ is a convex $L$-smooth function. However, when the constraint set $\mathcal{C}$ is strongly-convex and $\inf_{x \in \mathcal{C}} \|\nabla f(x)\| > 0$, Algorithm 1 enjoys a linear convergence rate [Levitin and Polyak, 1966, Demyanov and Rubinov, 1970]. Later on, Dunn [1979] showed that linear rates are maintained when the constraint set satisfies a condition subsuming local strong-convexity. Interestingly, these linear convergence regimes *do not* require the strong-convexity of $f$, *i.e.* the lower quadratic additional structure comes from the constraint set rather than from the function. When $x^*$ is in the interior of $\mathcal{C}$ and $f$ is strongly convex, Algorithm 1 also enjoys a linear convergence rate [Guélat and Marcotte, 1986].

These two linear convergence regimes can both become arbitrarily bad as $x^*$ gets close to the border of $\mathcal{C}$, and do not apply in the limit case where the unconstrained optimum of $f$ lies at the boundary of $\mathcal{C}$. In this scenario, when the constraint set is strongly convex, Garber and Hazan [2015] prove a general sublinear rate of $\mathcal{O}(1/T^2)$ when $f$ is $L$-smooth and $\mu$-strongly convex. In early iterations, these convergence rates can beat badly-conditioned linear rates.

Other structural assumptions are known to lead to accelerated convergence rates. However, these require elaborate algorithmic enhancements of the original Frank-Wolfe algorithm. Polytopes received much attention in particular, with *corrective* or *away* algorithmic mechanisms [Guélat and Marcotte, 1986, Hearn et al., 1987] that lead to linear convergence rates under appropriate structures of the objective function [Garber and Hazan, 2013a, Lacoste-Julien and Jaggi, 2013, 2015, Beck and Shtern, 2017, Gutman and Pena, 2018, Pena and Rodriguez, 2018, Diakonikolas et al., 2020, Carderera et al., 2021]. Accelerated versions of Frank-Wolfe, when the constraint set is a trace-norm ball (a.k.a. nuclear balls) – which are neither polyhedral nor strongly convex [So, 1990] – have also received a lot of attention [Freund et al., 2017, Allen-Zhu et al., 2017, Garber et al., 2018] and are especially useful in matrix completion [Shalev-Shwartz et al., 2011, Harchaoui et al., 2012, Dudik et al., 2012].

**Contributions.** We show **(1)** accelerated convergence rates for the Frank-Wolfe algorithm when the constraint set is uniformly convex, generalizing the rates of [Polyak, 1966, Demyanov and Rubinov, 1970, Garber and Hazan, 2015]. This fills the gap between all known convergence rates, *i.e.* between $\mathcal{O}(1/T)$ and the linear rate of [Levitin and Polyak, 1966, Demyanov and Rubinov, 1970, Dunn, 1979], and between $\mathcal{O}(1/T)$ and the $\mathcal{O}(1/T^2)$ rate of [Garber and Hazan, 2015] (see, *e.g.*, concluding remarks of [Garber and Hazan, 2015]). We also show **(2)** that accelerated convergence rates hold when the constraint set is only locally uniformly (or weaker) convex around the optimal solution thus explaining faster convergence rates (see Section 5). We state our convergence results **(3)** with generic structural assumptions, *e.g.*, Hölderian Error Bounds on $f$ replace usual strong convexity assumption. We also motivate (local) scaling inequalities as generic structural assumptions for (local) uniform convexity of the constraint set. Finally, we provide **(4)** similar arguments that interpolate between known regret bounds in an example of projection-free online learning. Namely, we prove accelerated regret bounds of the simple Follow-The-Leader (FTL) in online linear learning when the action set is uniformly convex and not necessarily smooth, see also [Huang et al., 2017, Molinaro, 2020]. Overall, we illustrate another key aspect of some projection-free algorithms: they are adaptive to many generic structural assumptions [Kerdreux, 2020].

**Outline.** In Section 2, we analyze the complexity of the Frank-Wolfe algorithm when the constraint set is uniformly convex, under various assumptions on $f$. In Section 2.3, we also establish accelerated convergence rate under weaker assumptions than global or local uniform convexity of the constraint set, see Section 5 for an illustration. In Section 3, we focus on the online linear optimization setting and provide analogous results to the previous section in terms of regret bounds for Follow-The-Leader (FTL). In Section 4, we give some examples of uniformly convex sets and relate the uniform convexity notion for sets with that of spaces and functions.

**Notation.** We use $d$ for the *ambient dimension* of the compact convex sets $\mathcal{C}$. We denote the *boundary* of $\mathcal{C}$ by $\partial\mathcal{C}$ and let $N_{\mathcal{C}}(x) \triangleq \{d \mid \langle d,\, y - x\rangle \leq 0,\ \forall y \in \mathcal{C}\}$ be the *normal cone* at $x$ with respect to $\mathcal{C}$. In the following, $x^*$ is an (optimal) solution to (OPT) and $(\alpha, q)$ denotes the uniform convexity parameters of a set. $p$ stands for the parameters for the various norm balls and might differ from $q$. We sometimes assume strict convexity of $f$ for the sake of exposition (only). Given a norm $\|\cdot\|$ we denote by $\|d\|_* \triangleq \max_{\|x\|\leq 1}\langle x,\, d\rangle$ its dual norm and we write the primal gap $h_t \triangleq f(x_t) - f(x^*)$. Finally, recall that a function is $L$-smooth on $\mathcal{C}$ w.r.t a norm $\|\cdot\|$ iff for any $(x, y) \in \mathcal{C}$

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

## 2 Frank-Wolfe Convergence Analysis with Uniformly Convex Constraints

In Theorem 2.2, we show accelerated convergence rate of the Frank-Wolfe algorithm when the constraint set $\mathcal{C}$ is $(\alpha, q)$-uniformly convex (with $q \geq 2$) and the smooth convex function satisfies $\inf_{x\in\mathcal{C}}\|\nabla f(x)\| > 0$; this is the interesting case. In Section 2.3, we then explore *localized* uniform convexity on the set $\mathcal{C}$ and provide convergence rates in Theorem 2.5. In Theorem 2.9 we show that $(\alpha, q)$-uniform convexity ensures convergence rates of the Frank-Wolfe algorithms in between the $\mathcal{O}(1/T)$ and $\mathcal{O}(1/T^2)$ [Garber and Hazan, 2015] when the function is strongly convex (and $L$-smooth), or satisfies a quadratic error bound at $x^*$. We also provide generalized convergence rates assuming Hölderian Error Bounds on $f$. In all of these scenarios, when the set is uniformly convex, the Frank-Wolfe algorithm (with short step as in Line 3 of Algorithm 1) enjoys accelerated convergence rates with respect to $\mathcal{O}(1/T)$.

**Proof Sketch.** We now provide an informal discussion as to why the uniform convexity of $\mathcal{C}$ leads to accelerated convergence rates under the classical assumptions that $\inf_{x\in\mathcal{C}}\|\nabla f(x)\| > 0$, *i.e.* $x^* \in \partial\mathcal{C}$.

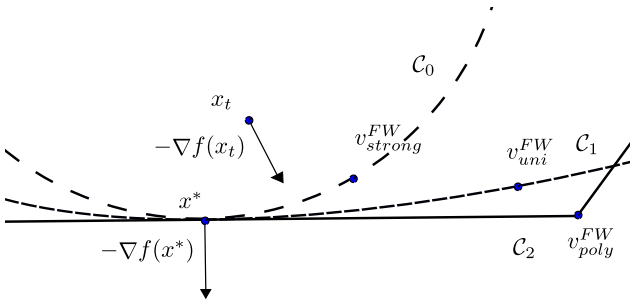


Figure 1: $v^{FW}_{strong}$, $v^{FW}_{uni}$, $v^{FW}_{poly}$ represent the various FW vertices from the strongly convex set $\mathcal{C}_0$, the uniformly convex set $\mathcal{C}_1$ and the polytope $\mathcal{C}_2$.

Formal arguments are developed in the proof of Theorem 2.2. The key point is that if $\mathcal{C}$ is curved around $x^*$ and $f$ is $L$-smooth, when $\|x_t - x^*\|$ converges to zero, the quantity $\|x_t - v_t\|$ also converges to zero, which is generally not the case, for instance when the constraint set is a polytope.

In Figure 1 we show various such behaviors. Applying the $L$-smoothness of $f$ to the Frank-Wolfe iterates, the classical iteration inequality is of the form (with $\gamma \in [0, 1]$)

$$h_{t+1} \leq h_t - \gamma\langle -\nabla f(x_t),\, v_t - x_t\rangle + \frac{\gamma^2}{2}L\|x_t - v_t\|^2. \quad (1)$$

The non-negative quantity $\langle -\nabla f(x_t),\, v_t - x_t\rangle$ participates in guaranteeing the function decrease, counterbalanced with $\|x_t - v_t\|^2$. The convergence rate then depends on specific relative quantification of these various terms, that we call *scaling inequalities* in Lemma 2.1 (which is a known equivalent definition of uniform convexity [Deville et al., 1993]) and Lemma 2.4.

### 2.1 Scaling Inequality on Uniformly Convex Sets

The following lemma outlines that the uniform convexity of $\mathcal{C}$ implies an upper bound on the distance between the current iterate and the Frank-Wolfe vertex as a power of the Frank-Wolfe gap $g(x_t) \triangleq \langle \nabla f(x_t),\, x_t - v_t\rangle$.

**Lemma 2.1** (Scaling Inequality)**.** *Assume the compact $\mathcal{C} \subset \mathbb{R}^d$ is an $(\alpha, q)$-uniformly convex set with respect to a norm $\|\cdot\|$, with $\alpha > 0$ and $q \geq 2$. Consider $x \in \mathcal{C}$, $\phi \in \mathbb{R}^d$ and $v_\phi \in \operatorname{argmax}_{v\in\mathcal{C}}\langle \phi, v\rangle$. Then, we have $\langle \phi, v_\phi - x\rangle \geq \frac{\alpha}{2}\|v_\phi - x\|^q\|\phi\|_*$. In particular for an iterate $x_t$ and its associated Frank-Wolfe vertex $v_t$, this yields*

$$\langle -\nabla f(x_t),\, v_t - x_t\rangle \geq \frac{\alpha}{2}\|v_t - x_t\|^q\|\nabla f(x_t)\|_*. \quad (2)$$

*Proof.* Because $\mathcal{C}$ is $(\alpha, q)$-uniformly convex, we have that for any $z \in \mathbb{R}^d$ of unit norm $(x+v_\phi)/2+\alpha/4\|x-v_\phi\|^q z \in \mathcal{C}$. By optimality of $v_\phi$, we have $\langle \phi, v_\phi\rangle \geq \langle \phi, (x+v_\phi)/2\rangle + \alpha/4\|x-v_\phi\|^q\langle \phi, z\rangle$. Hence, choosing the best $z$ implies $\langle \phi, v_\phi - x\rangle \geq \alpha/2\|v_\phi - x\|^q\|\phi\|_*$. ■

In other words, when $\mathcal{C}$ is uniformly convex, (2) quantifies the trade-off between the Frank-Wolfe gap $g(x_t) = \langle \nabla f(x_t), x_t - v_t\rangle$ and the value of $\|x_t - v_t\|$ under consideration in (1).

### 2.2 Interpolating Linear and Sublinear Rates

To our knowledge, no accelerated convergence rate of the Frank-Wolfe algorithm is known when the constraint set is uniformly convex but not strongly convex. We fill this gap in Theorem 2.2 below. In Section 4.2, we note that for $q \geq 2$, the $\ell_q$ are $(1/q, q)$-uniformly convex. Hence, at least for this specific family, when $q$ goes to $+\infty$, we can recover the classic sublinear convergence rate of $\mathcal{O}(1/T)$. For general families of $(\alpha, q)$-uniformly convex set, we do not know how the $\alpha$ parameter evolves with the uniform convexity exponent $q$.

**Theorem 2.2.** *Consider a convex $L$-smooth function $f$ and a compact convex set $\mathcal{C}$. Assume that $\mathcal{C}$ is $(\alpha, q)$-uniformly convex set with respect to a norm $\|\cdot\|$, with $q \geq 2$ and $\alpha > 0$. Assume $\|\nabla f(x)\|_* \geq c > 0$ for all $x \in \mathcal{C}$. Then the iterates of the Frank-Wolfe algorithm, with short step as in Line 3 of Algorithm 1 or exact line search, satisfy*

$$\begin{cases} h_T \leq M/(T+k)^{1/(1-2/q)} & \text{when } q > 2 \\ h_T \leq \big(1-\rho\big)^T h_0 & \text{when } q = 2, \end{cases}$$

*with $\rho = max\big\{\frac{1}{2}, 1 - c\alpha/L\big\}$, $k \triangleq (2-2^\eta)/(2^\eta - 1)$ and $M \triangleq max\{h_0 k^{1/\eta}, 2/((\eta-(1-\eta)(2^\eta-1))C)^{1/\eta}\}$, where $\eta \triangleq 1 - 2/q$ and $C \triangleq (c\alpha/2)^{2/q}/(2L)$.*

*Proof.* By $L$-smoothness of $f$ and because of the short step, we have for $\gamma \in [0, 1]$

$$f(x_{t+1}) \leq f(x_t) - \gamma g(x_t) + \frac{\gamma^2}{2} L\|x_t - v_t\|^2,$$

where $g(x_t)$ is the Frank-Wolfe gap.

With $\gamma = \min\big\{1, g(x_t)/(L\|x_t - v_t\|^2)\big\}$ we have

$$f(x_{t+1}) \leq f(x_t) - \frac{g(x_t)}{2} \cdot \min\Big\{1; \frac{g(x_t)}{L\|x_t - v_t\|^2}\Big\}.$$

Applying Lemma 2.1 with $\phi = -\nabla f(x_t)$ gives $g(x_t) \geq \alpha/2\|x_t - v_t\|^q\|\nabla f(x_t)\|_*$. Then

$$\begin{aligned} \frac{g(x_t)}{\|x_t - v_t\|^2} &= \Big(\frac{g(x_t)^{q/2-1} g(x_t)}{\|x_t - v_t\|^q}\Big)^{2/q} \\ &\geq \Big(\alpha/2\|\nabla f(x_t)\|_*\Big)^{2/q} g(x_t)^{1-2/q}. \end{aligned}$$

Finally, because $g(x_t) \geq f(x_t) - f(x^*) = h_t$, we have

$$h_{t+1} \leq h_t - \frac{h_t}{2}\min\Big\{1; \Big(\frac{\alpha}{2}\|\nabla f(x_t)\|_*\Big)^{\frac{2}{q}} h_t^{1-\frac{2}{q}}/L\Big\},$$

and finally,

$$h_{t+1} \leq h_t \cdot \max\Big\{\frac{1}{2}; 1 - \Big(\frac{\alpha}{2}\|\nabla f(x_t)\|_*\Big)^{\frac{2}{q}} h_t^{1-\frac{2}{q}}/(2L)\Big\}.$$

Then, by assumption, for all $x \in \mathcal{C}$, we have $\|\nabla f(x)\|_* > c > 0$ and hence we obtain

$$h_{t+1} \leq h_t \cdot \max\Big\{\frac{1}{2}; 1 - \Big(\frac{c\alpha}{2}\Big)^{\frac{2}{q}} \frac{h_t^{1-\frac{2}{q}}}{2L}\Big\}.$$

We solve the recursion with Lemma A.1; when $q = 2$ we recover the linear convergence rate. ■

**Remark 2.3.** *The convergence rates in Theorem 2.2 imply convergence rates in terms of distance to optimum by applying Lemma 2.1 with $\phi = -\nabla f(x^*)$ and convexity of $f$. Indeed, this yields*

$$\|x_t - x^*\|^q \leq \frac{2}{c\alpha}\langle -\nabla f(x^*), x^* - x_t\rangle \leq \frac{2}{c\alpha}\big(f(x_t) - f(x^*)\big).$$

*Hence, to obtain convergence rates in terms of the distance of the iterates to the optimum, the uniform convexity of the set supersedes that of the function, which is not needed here.*

### 2.3 Convergence Rates with Local Uniform Convexity

Theorem 2.2 relies on the global uniform convexity of the set. Actually, for the strongly convex case, it is equivalent to the global scaling inequality (2), see, *e.g.*, [Goncharov and Ivanov, 2017, Theorem 2.1 (g)]. However, weaker assumptions also lead to accelerated convergence rates of the Frank-Wolfe algorithm. In Theorem 2.5, we show accelerated convergence rates assuming a *local* scaling inequality at $x^*$. We then study the sets for which such an inequality holds. We say that a *local scaling inequality* holds at $x^* \in \mathcal{C}$, when there exists an $\alpha > 0$ and $q \geq 2$ such that for all $x \in \mathcal{C}$

$$\langle -\nabla f(x^*), x^* - x\rangle \geq \alpha/2\|\nabla f(x^*)\|_* \cdot \|x^* - x\|^q. \quad (3)$$

This combines the position of $-\nabla f(x^*)$ with respect to the normal cone of $\mathcal{C}$ at $x^*$ and the local geometry of $\mathcal{C}$ at $x^*$, see Remark 2.7. When the set $\mathcal{C}$ is globally $(\alpha, q)$-uniformly convex, this is a direct consequence of Lemma 2.1 because $-\nabla f(x^*) \in N_{\mathcal{C}}(x^*)$. In the following lemma, we prove that it is also a consequence of a natural definition of local uniform convexity of $\mathcal{C}$ at $x^*$. A proof is given in Appendix A.1.

**Lemma 2.4.** *Consider a compact convex set $\mathcal{C}$ and $x^*$ a solution to* (OPT). *Assume that $\mathcal{C}$ is locally $(\alpha, q)$-uniformly convex at $x^*$ with respect to $\|\cdot\|$ in the sense that, for all $x \in \mathcal{C}$, $\eta \in [0,1]$ and unit norm $z \in \mathbb{R}^d$, we have $\eta x^* + (1-\eta)x + \eta(1-\eta)\alpha\|x^* - x\|^q z \in \mathcal{C}$. Then* (3) *holds at $x^*$ with parameters $(\alpha, q)$.*

We obtain sublinear convergence rates that are systematically better than the $\mathcal{O}(1/T)$ baseline for any $q \geq 2$. A proof is deferred to Appendix A.3. In Section 5 we illustrate the benefice of local analysis.

**Theorem 2.5.** *Consider $f$ an $L$-smooth convex function and a compact convex set $\mathcal{C}$. Assume $\|\nabla f(x)\|_* > c > 0$ for all $x \in \mathcal{C}$ and write $x^* \in \partial\mathcal{C}$ a solution of* (OPT). *Further, assume that the convex set $\mathcal{C}$ satisfies a local scaling inequality at $x^*$ with parameters $(\alpha, q)$. Then the iterates of the Frank-Wolfe algorithm, with short step satisfy*

$$\begin{cases} h_T \leq M/(T+k)^{\frac{1}{1-2/(q(q-1))}} & \text{when } q > 2 \\ h_T \leq \big(1-\rho\big)^T h_0 & \text{when } q = 2, \end{cases}$$

*with $\rho = max\big\{\frac{1}{2}, 1 - c\alpha/L\big\}$, $k \triangleq (2-2^\eta)/(2^\eta - 1)$ and $M \triangleq max\{h_0 k^{1/\eta}, 2/((\eta-(1-\eta)(2^\eta-1))C)^{1/\eta}\}$, where $\eta \triangleq 1 - 2/(q(q-1))$ and $C \triangleq 1/(2LH^2)$. Note that $H$ depends only on $C, \alpha, L$ and $q$ (see Lemma A.2).*

**Remark 2.6.** *When the local scaling inequality* (3) *holds with $q = 2$, we obtain the same linear convergence regime as in Theorem 2.2. With $q > 2$, the sublinear convergence rates are of order $\mathcal{O}(1/T^{1/(1-2/(q(q-1)))})$ instead of $\mathcal{O}(1/T^{1/(1-2/q)})$ when the set is $(\alpha, q)$-uniformly convex and the global scaling inequality* (2) *holds. It is an open question to close this gap in the convergence regime with the local scaling inequality only.*

A similar approach appears in [Dunn, 1979] which introduces the following functional

$$a_{x^*}(\sigma) \triangleq \inf_{\substack{x \in \mathcal{C} \\ \|x - x^*\| \geq \sigma}} \langle \nabla f(x^*), x - x^* \rangle,$$

and shows that when there exists $A > 0$ such that $a_{x^*}(\sigma) \geq A\|x - x^*\|^2$, then the Frank-Wolfe algorithm converges linearly, under appropriate line-search rules. This result of [Dunn, 1979] thus subsumes that of [Levitin and Polyak, 1966, Demyanov and Rubinov, 1970]. However, no analysis was conducted for uniformly (but not strongly) convex set.

In Lemma 2.4 we showed that a given quantification of local uniform convexity implies the local scaling inequality and hence accelerated convergence rates. However, there are many situations where such a local notion of uniform convexity does not hold but (3) does. This was the essence of [Dunn, 1979, Remark 3.5.] that we state here. A proof is given in Appendix A.1.

**Corollary 2.7.** *Assume there exists a compact and $(\alpha, q)$-uniformly convex set $\Gamma$ such that $\mathcal{C} \subset \Gamma$ and $N_\Gamma(x^*) \subset N_\mathcal{C}(x^*)$, where $x^*$ is the solution of* (OPT). *If $-\nabla f(x^*) \in N_\Gamma(x^*)$, then* (3) *holds at $x^*$ with the $(\alpha, q)$ parameters.*

There exist numerous notions of local uniform convexity of a set that may imply local scaling inequalities. See for instance, the local directional strong convexity in [Goncharov and Ivanov, 2017, §Local Strong Convexity]. Alternatively, in the context of functions, Hölderian Errors Bounds (HEB) offer a weaker description of localized uniform convexity assumptions while retaining the same convergence rates [Kerdreux et al., 2019]. And these are known to hold generically for various classes of function [Lojasiewicz, 1965, Kurdyka, 1998, Bolte et al., 2007]. Obtaining a similar characterization for set is of interest. In particular, it is natural to relate enhanced convexity properties of the set *gauge function* $\|\cdot\|_\mathcal{C}$ [Rockafellar, 1970, §15] to convexity properties of the set or directly to local scaling inequalities. Error bounds as guaranteed with Łojasiewicz-type arguments on the gauge function could imply local scaling inequalities, showing that theses inequalities hold somewhat generically. A precise treatment of these questions is however out of the scope of this paper, see, *e.g.*, [Kerdreux et al., 2021] for some connection between the properties of the set gauge function and the set uniform convexity.

### 2.4 Interpolating Sublinear Rates for Arbitrary $x^*$

When the function is $\mu$-strongly convex and the set $\mathcal{C}$ is $\alpha$-strongly convex, Garber and Hazan [2015] show that the Frank-Wolfe algorithm (with short step) enjoys a general $\mathcal{O}(1/T^2)$ convergence rate. In particular, this result does not depend on the location of $x^*$ with respect to $\mathcal{C}$. We now generalize this result by relaxing the strong convexity of the constraint set $\mathcal{C}$ and the quadratic error bound on $f$.

**Hölderian Error Bounds.** Let $f$ be a strictly convex $L$-smooth function and $x^* = \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ where $\mathcal{C}$ is a compact convex set; the strict convexity assumption is only required to simplify exposition and the results hold more generally with the usual generalizations. We say that $f$ satisfies a $(\mu, \theta)$-Hölderian Error Bound when there exists $\theta \in [0, 1/2]$ such that

$$\|x - x^*\| \leq \mu(f(x) - f(x^*))^\theta. \tag{HEB}$$

When the function $f$ is subanalytic, (HEB) is known to hold generically [Lojasiewicz, 1965, Kurdyka, 1998, Bolte et al., 2007]. For instance, when $f$ is $(\mu, r)$-uniformly convex with $r \geq 2$ (see Definition C.1), then

it satisfies a $((2/\mu)^{1/r}, 1/r)$-Hölderian Error Bound, which follows from

$$f(x_t) \geq f(x^*) + \underbrace{\langle \nabla f(x^*),\, x_t - x^* \rangle}_{\geq 0} + \frac{\mu}{2}\|x_t - x^*\|_2^r.$$

Hence, we generalize the convergence result of [Garber and Hazan, 2015] and show that as soon as the set $\mathcal{C}$ is $(\alpha, q)$-uniformly convex with $q \geq 2$ and the function $f$ satisfies a non-trivial $(\mu, \theta)$-HEB, the Frank-Wolfe algorithm (with short step) enjoys an accelerated convergence rate with respect to $\mathcal{O}(1/T)$. In particular when $f$ is $\mu$-strongly convex, it satisfies a $(\mu, 1/2)$-HEB and by varying $q \geq 2$ we interpolate all sublinear convergence rates between $\mathcal{O}(1/T)$ and $\mathcal{O}(1/T^2)$.

In Lemma 2.8, we show an upper bound on $\|x_t - v_t\|$ when combining the uniform convexity of $\mathcal{C}$ and a Hölderian Error Bound for $f$. Lemma 2.8 is then the basis for the convergence analysis and similar to Lemma 2.1. The proofs are deferred to Appendix A.4.

**Lemma 2.8.** *Consider a compact and $(\alpha, q)$-uniformly convex set $\mathcal{C}$ with respect to $\|\cdot\|$. Denote $f$ a strictly convex $L$-smooth function and $x^* = \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Assume that $f$ satisfies a $(\mu, \theta)$-HEB $\|x - x^*\| \leq \mu(f(x) - f(x^*))^\theta$ with $\theta \in [0, 1/2]$. Then for $x_t \in \mathcal{C}$ we have $\alpha/(2\mu)\|x_t - v_t\|^q h_t^{1-\theta} \leq g(x_t)$, where $g(x_t)$ is the Frank-Wolfe gap and $v_t$ the Frank-Wolfe vertex.*

**Theorem 2.9.** *Consider a $L$-smooth convex function $f$ that satisfies a $(\mu, \theta)$-HEB with $\mu > 0$ and $\theta \in ]0, 1/2]$. Assume $\mathcal{C}$ is a compact and $(\alpha, q)$-uniformly convex set with respect to $\|\cdot\|$ with $q \geq 2$. Then the iterates of the Frank-Wolfe algorithm, with short step or exact line search, satisfy*

$$h_T \leq M/(T + k)^{1/(1-2\theta/q)},$$

*with $k \triangleq (2 - 2^\eta)/(2^\eta - 1)$ and $M \triangleq \max\{h_0 k^{1/\eta}, 2/((\eta - (1-\eta)(2^\eta - 1))C)^{1/\eta}\}$, where $\eta \triangleq 1 - 2\theta/q$ and $C \triangleq (\alpha/(2\mu))^{2/q}/L$. In particular for $q = 2$ and $\theta = 1/2$, we obtain the $\mathcal{O}(1/T^2)$ of [Garber and Hazan, 2015].*

Overall, Theorem 2.2, Theorem 2.5 and Theorem 2.9 give an (almost) complete picture of all the accelerated convergence regimes one can expect with the vanilla Frank-Wolfe algorithm.

## 3 Online Learning with Linear Oracles and Uniform Convexity

In online convex optimization, the algorithm sequentially decides an action, a point $x_t$ in a set $\mathcal{C}$, and then incurs a (convex smooth) loss $l_t(x_t)$. Algorithms are designed to reduce the cumulative incurred losses over time, $F_t = \frac{1}{t}\sum_{\tau=1}^t l_\tau(x_\tau)$. The comparison to the best action in hindsight is then defined as the *regret* of the algorithm, *i.e.* $R_T \triangleq \sum_{t=1}^T l_t(x_t) - \min_{x\in\mathcal{C}} \sum_{t=1}^T l_t(x)$.

Interesting correspondences have been established between the Frank-Wolfe algorithm and online learning algorithms. For instance, recent works [Abernethy and Wang, 2017, Abernethy et al., 2018] derive new Frank-Wolfe-like algorithms and analyses via two online learning algorithms playing against each other. Furthermore, a series of work proposed projection-free online algorithms inspired by their offline counterpart, *e.g.*, Hazan and Kale [2012] design a Frank-Wolfe online algorithm. In following works, Garber and Hazan [2013a,b] propose projection-free algorithms for online and offline optimization with optimal convergence guarantees where the decision sets are polytopes and the loss functions are strongly-convex. In the same setting, Lafond et al. [2015] analyze the online equivalent of the away-step Frank-Wolfe algorithm via a similar analysis to [Lacoste-Julien and Jaggi, 2013, 2015] in the offline setting. Recently, Hazan and Minasyan [2020] proposed a randomized projection-free algorithm that has a regret of $\mathcal{O}(T^{2/3})$ with high probability improving over the deterministic $\mathcal{O}(T^{3/4})$ of [Hazan and Kale, 2012] and Levy and Krause [2019] designed a projection-free online algorithm over smooth decision sets; dual to uniformly convex sets [Vial, 1983].

**Online Linear Optimization and Set Curvature.** At a high level, when the constraint set is strongly-convex, the analyses of the simple Follow-The-Leader (FTL) for online linear optimization [Huang et al., 2016] is analogous to the offline convergence analyses of the Frank-Wolfe algorithm when not assuming strong-convexity of the objective function as in [Polyak, 1966, Demyanov and Rubinov, 1970, Dunn, 1979]. Indeed, by definition, linear functions do not enjoy non-linear lower bounds, *i.e.* uniform convexity-like assumptions.

In the online linear setting, we write the functions $l_t(x) = \langle c_t, x\rangle$ and assume that $(c_t)$ belong to a bounded set $\mathcal{W}$ (smoothness). FTL consists in choosing the action $x_t$ at time $t$ that minimizes the cumulative sum of the previously observed losses, *i.e.* each iteration solves the minimization of a linear function over $\mathcal{C}$

$$x_T \in \operatorname*{argmin}_{x\in\mathcal{C}} \sum_{t=1}^{T-1} l_t(x) = \langle \sum_{t=1}^{T-1} c_t,\, x\rangle. \tag{4}$$

In general, FTL incurs a worst-case regret of $\mathcal{O}(T)$ [Shalev-Shwartz et al., 2012]. For online linear learning, Huang et al. [2016, 2017] study the conditions

under which the strong convexity of the decision set $\mathcal{C}$ leads to improved regret bounds. In particular, when there exists a $c > 0$ such that for all $T$, $\min_{1\leq t\leq T}\|\frac{1}{t}\sum_{\tau=1}^{t} c_\tau\|_* \geq c > 0$, then FTL enjoys the optimal regret bound of $\mathcal{O}(\log(T))$ [Huang et al., 2017]. Molinaro [2020] extends this result by dropping the smoothness assumption on $\mathcal{C}$ required in [Huang et al., 2017]. This convergence result with FTL is the counter part of the offline linear convergence analyses of the Frank-Wolfe algorithm when $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_* \geq c > 0$ and $\mathcal{C}$ is a strongly convex set [Polyak, 1966, Demyanov and Rubinov, 1970, Dunn, 1979]. In Theorem 3.1, we hence further support this analogy between online and offline settings. We show that FTL enjoys continuously interpolated regret bounds between $\mathcal{O}(\log(T))$ and $\mathcal{O}(T)$ for all types of uniform convexity of the decision sets. Again, this covers a much broader spectrum of *curved* sets, and is similar to Theorem 2.2 in the Frank-Wolfe setting. A proof is deferred to Appendix B.

**Theorem 3.1.** *Let $\mathcal{C}$ be a compact and $(\alpha, q)$-uniformly convex set with respect to $\|\cdot\|$. Assume that $L_T = min_{1\leq t\leq T}\|\frac{1}{t}\sum_{\tau=1}^{t} c_\tau\|_* > 0$. Then the regret $R_T$ of FTL (4) for online linear optimization satisfies*

$$\begin{cases} R_T \leq 2M\Big(\dfrac{2M}{\alpha L_T}\Big)^{\frac{1}{q-1}}\Big(\dfrac{q-1}{q-2}\Big)T^{1-\frac{1}{q-1}}, & q > 2 \\ R_T \leq \dfrac{4M^2}{\alpha L_T}(1+\log(T)), & q = 2, \end{cases}$$

*where $M = sup_{c\in\mathcal{W}}\|c\|_*$, with the losses $l_t(x) = \langle c_t, x\rangle$ and $(c_t)$ belong to the bounded set $\mathcal{W}$.*

With the simple FTL, we obtain non-trivial regret bounds, *i.e.*, $o(T)$, whenever the set is uniformly convex, without any curvature assumption on the loss functions (because they are linear). In particular for $q \in [2, 3]$, it improves over the general tight regret bound of $\mathcal{O}(\sqrt{T})$ for smooth convex losses and compact convex decision sets [Shalev-Shwartz et al., 2012]. Interestingly, with the same assumption on $\mathcal{C}$, Dekel et al. [2017] obtain for online linear optimization, the same asymptotical regret bounds with a variation of Follow-The-Leader incorporating hints. It is remarkable that the presence of hints or the assumption $\min_{1\leq t\leq T}\|\frac{1}{t}\sum_{\tau=1}^{t} c_\tau\|_* \geq c > 0$ for all $T$ both lead to the same bounds.

# 4 Examples of Uniformly Convex Objects

The uniform convexity assumptions refine the convex properties of several mathematical objects, such as normed spaces, functions, and sets. In this section, we provide some connection between these various notions of uniform convexity, see, *e.g.*, [Kerdreux et al., 2021] for an in-depth discussion. In Section 4.1, we recall that norm balls of uniformly convex spaces are uniformly convex sets, and show set uniform convexity of classic norm balls in Section 4.2. In Appendix C.2, we show that the level sets of some uniformly convex functions are uniformly convex sets, extending the strong convexity results of [Garber and Hazan, 2015, Section 5].

## 4.1 Uniformly Convex Spaces

The uniform convexity of norm balls (Definition 1.1) is closely related to the uniform convexity of normed spaces [Polyak, 1966, Balashov and Repovs, 2011, Lindenstrauss and Tzafriri, 2013, Weber and Reisig, 2013]. Some classical works establish sharp uniform convexity results for classical normed spaces such as $l_p$, $L_p$ or $C_p$. Most of the practical examples of uniformly convex sets are norm balls and are hence tightly linked with uniformly convex spaces. The property of these sets has many consequences, *e.g.*, in learning theory [Donahue et al., 1997]. It also relates to concentration inequalities in Banach Spaces [Juditsky and Nemirovski, 2008] and hence implications [Ivanov, 2019] for approximate versions of the Carathéodory theorem [Combettes and Pokutta, 2019].

Clarkson [1936], Boas Jr [1940] define a uniformly convex normed space $(\mathbb{X}, \|\cdot\|)$ as a normed space such that, for each $\epsilon > 0$, there is a $\delta > 0$ such that if $x$ and $y$ are unit vectors in $\mathbb{X}$ with $\|x - y\| \geq \epsilon$, then $(x + y)/2$ has norm lesser or equal to $1 - \delta$. Specific quantification of spaces satisfying this property is obtained via the modulus of convexity, a measure of non-linearity of a norm.

**Definition 4.1** (Modulus of convexity)**.** *The modulus of convexity of the space $(\mathbb{X}, \|\cdot\|)$ is defined as*

$$\delta_{\mathbb{X}}(\epsilon) = \inf_{\|x\|,\|y\|\leq 1}\Big\{1 - \Big\|\frac{x+y}{2}\Big\| \;\Big|\; \|x-y\| \geq \epsilon\Big\}.$$

A normed space $\mathbb{X}$ is said to be $r$-uniformly convex in the case $\delta_{\mathbb{X}}(\epsilon) \geq C\epsilon^r$. These specific lower bounds on the modulus of convexity imply that the balls stemming for such spaces are uniformly convex in the sense of Definition 1.1. There exist sharp results for $L_p$ and $\ell_p$ spaces in [Clarkson, 1936, Hanner et al., 1956]. Matrix spaces with $p$-Schatten norm are known as $C_p$ spaces, and sharp results concerning their uniform convexity can be found in [Dixmier, 1953, Tomczak-Jaegermann, 1974, Simon, 2005, Ball et al., 1994]. The following gives a link between the set $\gamma_{\mathcal{C}}$ and space $\delta_{\mathbb{X}}$ modulus of convexity, see proof in Appendix C.1 or [Molinaro, 2020, Appendix A].

**Lemma 4.2.** *If a normed space $(\mathbb{X}, \|\cdot\|)$ is uniformly convex with modulus of convexity $\delta_{\mathbb{X}}(\cdot)$, then*

*its unit norm ball is $\delta_{\mathbb{X}}(\cdot)$ uniformly convex with respect to $\|\cdot\|$. Note that if the unit ball $B_{\|\cdot\|}(1)$ is $(\alpha, q)$-uniformly convex w.r.t. $\|\cdot\|$, then $B_{\|\cdot\|}(r)$ is $(\alpha/r^{q-1}, q)$-uniformly convex w.r.t. $\|\cdot\|$.*

### 4.2 Uniform Convexity of Some Classic Norm Balls

When $p \in ]1, 2]$, $\ell_p$-balls are strongly convex sets and $((p-1)/2, 2)$-uniformly convex with respect to $\|\cdot\|_p$, see, *e.g.*, [Hanner et al., 1956, Theorem 2] or [Garber and Hazan, 2015, Lemma 4]. When $p > 2$, the $\ell_p$-balls are $(1/p, p)$-uniformly convex with respect to $\|\cdot\|_p$ [Hanner et al., 1956, Theorem 2]. Uniform convexity also extends the strong convexity of group $\ell_{s,p}$-norms (with $1 < p, s \leq 2$) [Garber and Hazan, 2015, §5.3. and 5.4.] to the general case $p, s > 1$.

Dixmier [1953], Tomczak-Jaegermann [1974], Simon [2005], Ball et al. [1994] focus of the uniform convexity of the $(C_p, \|\cdot\|_{S(p)})$ spaces, *i.e.* spaces of matrix where the norm is the $\ell_p$-norm of a matrix singular values . Their unit balls are hence the $p$-Schatten balls. For $p \in ]1, 2]$, $p$-Schatten balls are $((p-1)/2, 2)$-uniformly convex with respect to $\|\cdot\|_{S(p)}$, see [Garber and Hazan, 2015, Lemma 6] and the sharp results of [Ball et al., 1994]. For the case $p > 2$, Dixmier [1953] showed that the $p$-Schatten balls are $(1/p, p)$-uniformly convex with respect to $\|\cdot\|_{S(p)}$, see also [Ball et al., 1994, §III].

## 5 A Numerical Illustration

Uniform convexity is a global assumption. Hence, in Theorem 2.2, we obtain sublinear convergence that do not depend on the specific location of the solution $x^* \in \partial\mathcal{C}$. However, some regions of $\mathcal{C}$ might be relatively more curved than others and hence empirically exhibit faster convergence rates. We show that local scaling inequality is an adequate quantification of local curvature around the optimum, and we proved accelerated convergence rates in Theorem 2.5. We now provide some examples of these observed accelerated regimes.

In Figure 2-3, we solve (OPT) with the Frank-Wolfe algorithm where $f$ is a quadratic with condition number 100 and the constraint sets are various $\ell_p$-balls of radius 5. We vary $p$ so that all balls are uniformly convex but not strongly-convex. We also change the approximate location of the optimum $x^*$ in the boundary of the $\ell_p$-balls.

Subfigures (2a), and (2b) are associated to an optimization problem where the solution $x^*$ of (OPT) is near the intersection of the $\ell_p$-balls and the half-line generated by $\sum_{i=1}^{d} e_i$ (where the $(e_i)$ is the canonical basis), *i.e.* in *curved* regions of the boundaries of the

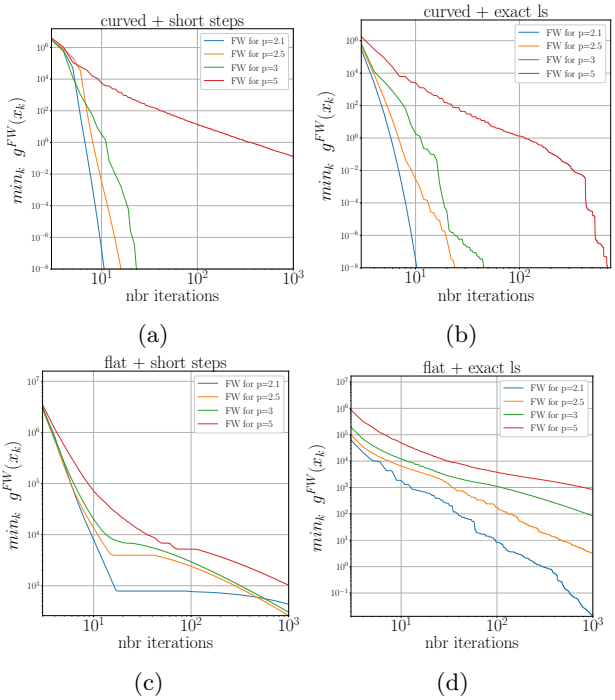


Figure 2: On a line, each plot exhibits the behavior of the Frank-Wolfe algorithm iterates with different step size strategy: deterministic line-search (*i.e.* $1/(k+1)$), short step and exact line-search. To avoid the oscillating behavior of Frank-Wolfe gap, the $y$-axis represents $\min_{k=1,\dots,T} g(x_k)$ where $g(\cdot)$ is the Frank-Wolfe gap and $T$ the number of iterations.

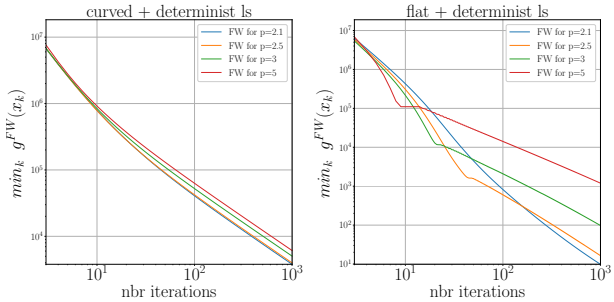


Figure 3: Same as Figure 2 but with deterministic line-search.

$\ell_p$-balls. Subfigures (2c), and (2d) corresponds to the same optimization problem where the solution $x^*$ to (OPT) is close to the intersection between the half-line generated by $e_1$ and the boundary of the $\ell_p$-balls, *i.e.* in *flat* regions of the boundaries of the $\ell_p$-balls. We observe that when the optimum is at a *curved* location, the convergence is quickly linear for $p$ sufficiently close to 2 and appropriate line-search (see Subfigures (2a) and (2b)). However, when the optimum is near the *flat* location, we indeed observe sublinear convergence rates (see Subfigures (2c) and (2d)). It still becomes linear for $p = 2.1$ with exact line-search in Subfigure (2d).

Also, Theorem 2.2 gives accelerated rates when using the Frank-Wolfe algorithm with exact line-search or short step. In Subfigures (3a) and (3b), we show examples of the convergence of the Frank-Wolfe algorithm when using deterministic line-search. The rates are indeed sublinear in $\mathcal{O}(1/T)$.

## 6 Conclusion

Our results fill the gap between known convergence rates for the Frank Wolfe algorithm and show that it is (also) the *curvature* of the constraint set that accelerates the convergence of the Frank-Wolfe algorithm, not just the strong convexity of the function or the set. In applications where the constraints are likely to be active (*e.g.*, regularization), the assumption that $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|_* \geq c > 0$ is not restrictive and the value of $c$ quantifies the relevance of the constraints.

Importantly, we also prove the linear convergence of the vanilla Frank-Wolfe without requiring the set's global strong convexity. Apart from [Dunn, 1979], we are not aware of any other work leveraging such local set properties as in Section 2.3. This could also have an impact in online learning or learning theory, see, *e.g.*, [Kerdreux et al., 2021, §6.2.] for a connection between the uniform convexity of a set and upper-bounds on Rademacher constants.

We provide similar results in online learning following the analysis of FTL on strongly convex sets [Huang et al., 2017, Molinaro, 2020]. We prove that the simple Follow-The-Leader (FTL) enjoys fast regret bounds when the action set is uniformly convex and without smoothness assumption on the domain. FTL and FW are both projection-free methods, *i.e.*, iterative algorithms minimizing a linear function on a given domain at each iteration.

This work also proves that projection-free methods can be adaptive to a variety of structural properties of (OPT), *i.e.* the unknown global or local uniform convexity parameters of the set or the unknown Hölderian Error Bounds (HEB) parameters of the function, *e.g.*, Theorem 2.9. These results hence complement other works emphasizing such properties, *e.g.*, the adaptive properties of *corrective* versions of Frank-Wolfe with HEB [Kerdreux et al., 2019] or the affine invariant analysis of FW in a variety of settings [Jaggi, 2013, Lacoste-Julien and Jaggi, 2013, Kerdreux et al., 2020].

In this paper, we also highlight the importance of the *scaling-inequalities* as useful characterizations of the uniform convexity of the sets as opposed to the classical definition. Finally, we note that in the infinite-dimensional settings, the set *curvature* (*i.e.* uniform convexity), or the situation where the optimum is in the interior of the constraints [Bach et al., 2012, Lacoste-Julien et al., 2015], are the *only* known structural sources of acceleration for the Frank-Wolfe algorithms.

## Acknowledgements

T.K. thanks Pierre-Cyril Aubin for very interesting discussions on Banach spaces, which contributed is the motivation for studying the convergence rates of projection-free methods with uniform convexity assumptions. T.K. also acknowledges funding from the CFM-ENS chaire *les modèles et sciences des données*. Research reported in this paper was 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+. AA is at the département d'informatique de l'École Normale Supérieure, UMR CNRS 8548, PSL Research University, 75005 Paris, France, and INRIA. AA would like to acknowledge support from the *ML and Optimisation* joint research initiative with the *fonds AXA pour la recherche* and Kamet Ventures, a Google focused award, as well as funding by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

## References

Abernethy, J., Lai, K. A., Levy, K. Y., and Wang, J.-K. (2018). Faster rates for convex-concave games. *arXiv preprint arXiv:1805.06792*.

Abernethy, J. D. and Wang, J.-K. (2017). On Frank-Wolfe and equilibrium computation. In *Advances in Neural Information Processing Systems*, pages 6584–6593.

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

Allen-Zhu, Z., Hazan, E., Hu, W., and Li, Y. (2017). Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls. In *Advances in Neural Information Processing Systems*, pages 6191–6200.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*.

Balashov, M. V. and Repovs, D. (2011). Uniformly convex subsets of the Hilbert space with modulus of convexity of the second order. *arXiv preprint arXiv:1101.5685*.

Ball, K., Carlen, E. A., and Lieb, E. H. (1994). Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482.

Beck, A. and Shtern, S. (2017). Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27.

Boas Jr, R. P. (1940). Some uniformly convex spaces. *Bulletin of the American Mathematical Society*, 46(4):304–311.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer.

Bojanowski, P., Lajugie, R., Grave, E., Bach, F., Laptev, I., Ponce, J., and Schmid, C. (2015). Weakly-supervised alignment of video with text. In *Proceedings of the IEEE international conference on computer vision*, pages 4462–4470.

Bolte, J., Daniilidis, A., and Lewis, A. (2007). The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223.

Canon, M. D. and Cullum, C. D. (1968). A tight upper bound on the rate of convergence of Frank-Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516.

Carderera, A., Diakonikolas, J., Lin, C. Y., and Pokutta, S. (2021). Parameter-free locally accelerated conditional gradients. *arXiv:2102.06806*.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40(3):396–414.

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

Combettes, C. W. and Pokutta, S. (2019). Revisiting the approximate Carathéodory problem via the Frank-Wolfe algorithm. *arXiv preprint arXiv:1911.04415*.

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

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

Dekel, O., Haghtalab, N., Jaillet, P., et al. (2017). Online learning with a hint. In *Advances in Neural Information Processing Systems*, pages 5299–5308.

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

Deville, R., Godefroy, G., and Zizler, V. (1993). *Smoothness and renormings in Banach spaces*. Longman Scientific Technical, Harlow.

Diakonikolas, J., Carderera, A., and Pokutta, S. (2020). Locally accelerated conditional gradients. In *International Conference on Artificial Intelligence and Statistics*, pages 1737–1747. PMLR.

Dixmier, J. (1953). Formes linéaires sur un anneau d'opérateurs. *Bulletin de la Société Mathématique de France*, 81:9–39.

Donahue, M. J., Darken, C., Gurvits, L., and Sontag, E. (1997). Rates of convex approximation in non-Hilbert spaces. *Constructive Approximation*, 13(2):187–220.

Dudik, M., Harchaoui, Z., and Malick, J. (2012). Lifted coordinate descent for learning with trace-norm regularization. In *Artificial Intelligence and Statistics*, pages 327–336.

Dunn, J. C. (1979). Rates of convergence for conditional gradient algorithms near singular and nonsingular ex-

tremals. *SIAM Journal on Control and Optimization*, 17(2):187–211.

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

Freund, R. M., Grigas, P., and Mazumder, R. (2017). An extended Frank–Wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346.

Futami, F., Cui, Z., Sato, I., and Sugiyama, M. (2019). Bayesian posterior approximation via greedy particle optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3606–3613.

Garber, D. and Hazan, E. (2013a). A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*.

Garber, D. and Hazan, E. (2013b). Playing non-linear games with linear oracles. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 420–428. IEEE.

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

Garber, D., Sabach, S., and Kaplan, A. (2018). Fast generalized conditional gradient method with applications to matrix recovery problems. *arXiv preprint arXiv:1802.05581*.

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

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

Gutman, D. H. and Pena, J. F. (2018). The condition of a function relative to a polytope. *arXiv preprint arXiv:1802.00271*.

Hanner, O. et al. (1956). On the uniform convexity of Lp and lp. *Arkiv för Matematik*, 3(3):239–244.

Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., and Malick, J. (2012). Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3393. IEEE.

Hazan, E. and Kale, S. (2012). Projection-free online learning. In *29th International Conference on Machine Learning, ICML 2012*, pages 521–528.

Hazan, E. and Minasyan, E. (2020). Faster projection-free online learning. In *Conference on Learning Theory*, pages 1877–1893. PMLR.

Hearn, D. W., Lawphongpanich, S., and Ventura, J. A. (1987). Restricted simplicial decomposition: Computation and extensions. In *Computation Mathematical Programming*, pages 99–118. Springer.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. (2016). Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems*, pages 4970–4978.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. (2017). Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *The Journal of Machine Learning Research*, 18(1):5325–5355.

Ivanov, G. (2019). Approximate Carathéodory's theorem in uniformly smooth Banach spaces. *Discrete & Computational Geometry*, pages 1–8.

Jaggi, M. (2011). Convex optimization without projection steps. *Arxiv preprint arXiv:1108.1170*.

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

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553.

Juditsky, A. and Nemirovski, A. S. (2008). Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*.

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

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

Kerdreux, T., d'Aspremont, A., and Pokutta, S. (2019). Restarting Frank-Wolfe. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1275–1283.

Kerdreux, T., Liu, L., Lacoste-Julien, S., and Scieur, D. (2020). Affine invariant analysis of Frank-Wolfe on strongly convex sets. *arXiv:2011.03351*.

Kurdyka, K. (1998). On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783.

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

Lacoste-Julien, S. and Jaggi, M. (2015). On the global linear convergence of Frank–Wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28:496–504.

Lacoste-Julien, S., Lindsten, F., and Bach, F. (2015). Sequential kernel herding: Frank-Wolfe optimization for particle filtering. *arXiv preprint arXiv:1501.02056*.

Lafond, J., Wai, H.-T., and Moulines, E. (2015). On the online Frank-Wolfe algorithms for convex and non-convex optimizations. *arXiv preprint arXiv:1510.01171*.

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

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

Levy, K. and Krause, A. (2019). Projection free online learning over smooth sets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1458–1466.

Lindenstrauss, J. and Tzafriri, L. (2013). *Classical Banach spaces II: function spaces*, volume 97. Springer Science & Business Media.

Lojasiewicz, S. (1965). Ensembles semi-analytiques. *Institut des Hautes Études Scientifiques*.

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

Miech, A., Alayrac, J.-B., Bojanowski, P., Laptev, I., and Sivic, J. (2017). Learning from video and text via large-scale discriminative clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5257–5266.

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

Nguyen, H. and Petrova, G. (2017). Greedy strategies for convex optimization. *Calcolo*, 54(1):207–224.

Paty, F.-P. and Cuturi, M. (2019). Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*.

Pena, J. and Rodriguez, D. (2018). Polytope conditioning and linear convergence of the Frank-Wolfe algorithm. *Mathematics of Operations Research*.

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

Polyak, B. T. (1966). Existence theorems and convergence of minimizing sequences for extremal problems with constraints. In *Doklady Akademii Nauk*, volume 166, pages 287–290. Russian Academy of Sciences.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press., Princeton.

Seguin, G., Bojanowski, P., Lajugie, R., and Laptev, I. (2016). Instance-level video segmentation from object tracks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3678–3687.

Shalev-Shwartz, S. (2007). *Online Learning: Theory, Algorithms, and Applications*. PhD thesis.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S., Gonen, A., and Shamir, O. (2011). Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*.

Simon, B. (2005). *Trace ideals and their applications*. American Mathematical Soc.

So, W. (1990). Facial structures of Schatten p-norms. *Linear and Multilinear Algebra*, 27(3):207–212.

Temlyakov, V. (2011). *Greedy approximation*, volume 20. Cambridge University Press.

Tomczak-Jaegermann, N. (1974). The moduli of smoothness and convexity and the rademacher averages of the trace classes. *Studia Mathematica*, 50(2):163–182.

Vial, J.-P. (1983). Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259.

Weber, A. and Reisig, G. (2013). Local characterization of strongly convex sets. *Journal of Mathematical Analysis and Applications*, 400(2):743–750.

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