Frank-Wolfe on Uniformly Convex Sets

Thomas Kerdreux\textsuperscript{1,3} and Alexandre d’Aspremont\textsuperscript{2,3}

\textsuperscript{1}INRIA, Paris
\textsuperscript{2}CNRS, UMR 8548
\textsuperscript{3}D.I., École Normale Supérieure, Paris, France

April 24, 2020

Abstract

The Frank-Wolfe method solves smooth constrained convex optimization problems at a generic sublinear rate of $O(1/T)$, and 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 induce accelerated convergence rates for the 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 for Frank-Wolfe. These results also importantly highlight that Frank-Wolfe is adaptive to much more generic constraint set structures, thus explaining faster empirical convergence.

1 Introduction

The Frank-Wolfe method (Frank and Wolfe, 1956) (Algorithm 1) is a projection-free algorithm that solves the generic optimization problem

$$\arg\min_{x \in C} f(x),$$

(1)

where $C$ 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 retained in the original Frank-Wolfe algorithm. Each iteration requires to solve a Linear Minimization Oracle (see line 2), instead of a projection or proximal operation that may not be computationally competitive in various settings. Also, the iterates of Algorithm 1 are convex combinations of extreme points of $C$, the solutions of the Linear Minimization Oracle (see line 2). Hence, depending on the extremal structure of $C$, the early iterates may have a specific structure, sparse or low rank for instance, that could be traded-off with the iterate approximation quality of problem (1).

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; Miech et al., 2018), quadrature rules in RKHS (Bach et al., 2012; Lacoste-Julien et al., 2015; Futami et al., 2018, 2019), optimal transport (Courty et al., 2016; Vayer et al., 2018; Paty and Cuturi, 2019; Luise et al., 2019) and many others.
Algorithm 1 Frank-Wolfe Algorithm

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

1: for $t = 0, 1, \ldots, T$ do
2: \hspace{1em} $v_t \in \arg\max_{v \in C} \langle -\nabla f(x_t); v - x_t \rangle$ \hfill \triangleright \text{Linear minimization oracle}
3: \hspace{1em} $\gamma_t = \arg\min_{\gamma \in [0, 1]} \gamma \langle v_t - x_t; \nabla f(x_t) \rangle + \frac{\eta^2}{2} L ||v_t - x_t||^2$ \hfill \triangleright \text{short step-size}
4: \hspace{1em} $x_{t+1} = (1 - \gamma_t)x_t + \gamma_t v_t$ \hfill \triangleright \text{Convex update}
5: end for

**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 $C$ is a compact convex set and $f$ a $L$-smooth function. When the constraint set $C$ is strongly-convex, the method can be accelerated with convergence rates depending on additional assumptions on $f$ and line-search type. In particular line 3 of Algorithm 1 simply corresponds to minimizing the quadratic upper bound on the function given by $L$-smoothness of $f$ at the current iterate. The results we state in Section 2 also hold with exact line-search or the adaptive scheme of (Pedregosa et al., 2018).

When the constraint set $C$ is strongly-convex and there exists $c > 0$ such that $||\nabla f(x^*)|| > c$, Algorithm 1 enjoys a linear convergence rate (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970). In particular, linear convergence does not require strong-convexity of the function $f$, i.e. the quadratic additional structure comes from the constraint set rather than from the function.

When $x^*$ is in the interior of $C$ 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 $C$, and do not apply in the limit case where the unconstrained optimum lies at the boundary of $C$. 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.

The work of (Dunn, 1979) significantly refined the type of constraint structures that induce accelerated convergence rates for the Frank-Wolfe algorithm. In particular, Dunn (1979) shows that the linear convergence rates for strongly-convex set (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970), can be obtained when the constraint set satisfies a condition subsuming local strong-convexity in non-trivial ways.

Here we prove new sublinear rates, when the constraint set is uniformly convex, for each of the two functional assumptions described above. In particular, these new rates continuously interpolate all the convergence regimes between the generic sublinear rate $\mathcal{O}(1/T)$ and the linear rate of (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970) on the one hand, and between the generic sublinear $\mathcal{O}(1/T)$ and the $\mathcal{O}(1/T^2)$ of (Garber and Hazan, 2015), depending on the location of the optimum.

Finally, other structural assumptions are known to lead to accelerated convergence rates. However, these require elaborate algorithmic enhancements on 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 $f$ (Garber and Hazan, 2013; Lacoste-Julien and Jaggi, 2013, 2015; Beck and Shtern, 2017; Gutman and Pena, 2018; Pena and Rodriguez, 2018). Accelerated versions of Frank-Wolfe, when the constraint set is a trace-norm ball (a.k.a. nuclear balls) – which are not 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 (Jaggi and Sulovský, 2010; Shalev-Shwartz et al., 2011; Harchaoui et al., 2012; Dudik et al., 2012).
Uniform convexity. Qualitatively, uniform convexity is a global quantification of the curvature of a convex set \( C \). There exist several definitions and we give one here generalizing the definition of strong convexity of a set.

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

\[
\eta x + (1 - \eta)y + \eta(1 - \eta)\gamma_C(||x - y||)z \in C,
\]

where \( \gamma_C(\cdot) \geq 0 \) is a non-decreasing function. In particular when there exists \( \alpha > 0 \) and \( q > 0 \) such that \( \gamma_C(r) = \alpha r^q \), we say that \( C \) is \((\alpha, q)\)-uniformly convex.

Intuitively, uniform convexity assumption strengthens the convexity property of \( C \) which means that any line segment between two points is included in \( C \) by requiring a scaled unit ball to fit in \( C \) and results in curved sets. Other characterizations of uniform convexity rely on this construction, formalized as lens (Blanc, 1943) or arc properties (Weber and Reisig, 2013). Strongly convex sets are uniformly convex sets for which \( \gamma_C(r) = \alpha r^2 \), i.e. \((\alpha, 2)\)-uniformly convex sets. Two common families of uniformly convex sets are the \( \ell_p \) balls and the \( p \)-Schatten balls for \( p \in [1; +\infty[ \), and from these we can also construct some uniformly convex group norm balls. As an example, \( \ell_p \) balls are uniformly convex for any \( p > 1 \) but strongly-convex for \( p \in [1, 2] \) only.

**Contributions.** We show accelerated sublinear converge rates for the Frank-Wolfe algorithm, with appropriate line-search, for smooth constrained optimization problems when the constraint set is uniformly convex. This generalizes the rates of (Polyak, 1966) and (Garber and Hazan, 2015) under their respective assumptions, and fills the gap between all known convergence rates (see e.g. concluding remarks of (Garber and Hazan, 2015)).

**Outline.** In Section 2, we provide analyse the complexity of Frank-Wolfe when the constraint set is uniformly convex, under various assumptions on \( f \). In Section 3, we give some examples of uniformly convex sets and relate the uniform convexity notion for sets with that applied to spaces and functions. We conclude in Section 4 with some numerical experiments.

**Notations.** We use \( d \) for the ambient dimension of the convex sets \( C \). \( K_C(x) \) denotes the normal cone at \( x \) with respect to \( C \). \( x^* \) stands for a solution to \((1)\). \((\alpha, q)\) (resp. \((\mu, r)\)) denote the uniform convexity parameters of the sets (resp. functions). \( p \) (resp. \( r \)) denotes the parameter (resp. radius) of norm balls and \( \mathbb{K} \) stands for a normed space, and we write \( e_i \) the canonical basis of \( \mathbb{R}^d \). We write the primal gap \( h_t = f(x_t) - f(x^*) \). For \( x \in \mathbb{R}^d \), \( ||x||_p = \left( \sum_{i=1}^d |x_i|^p \right)^{1/p} \). Finally for a matrix \( M \in \mathbb{R}^{d \times d} \), we write its \( p \)-Schatten norm as \( \|M\|_{S(p)} = \left( \sum_{i=1}^d \sigma_i(M)^p \right)^{1/p} \).

2 Analysis of Frank-Wolfe with uniform convexity on constraints

Previous works (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970; Dunn, 1979) showed that when \( C \) is strongly convex, \( ||\nabla f(x^*)|| > c > 0 \), and the function is \( L \)-smooth, then the Frank-Wolfe algorithm converges linearly, under specific line-search strategies such as short step size, exact line-search or some specialized variants (Dunn, 1979). Hence, there is a gap between the general sublinear rate of \( O(1/T) \) for compact convex sets and this linear convergence for strongly convex sets. In Theorem 2.1, we show that \((\alpha, q)\) uniform convexity (with \( q \geq 2 \)) of the constraint set \( C \) leads to a continuous interpolation of these...
two convergence regimes. Importantly, the value of $q$ measuring the uniform convexity is known for many classical sets as detailed in Section 3.

Assumption $||\nabla f(x^*)|| > 0$ implies that the unconstrained optimum of $f$ does not belong to $C$. When the unconstrained optimum of $f$ is in the interior of the set $C$, there is also linear convergence of the Frank-Wolfe iterates (again with the short-step sizes) (Guélat and Marcotte, 1986). These two results do not account for the unlikely case where unconstrained optimum exactly lies on the boundary of $C$. Also, the convergence rate depends on the norm of $||\nabla f(x^*)||$ or on the distance of $x^*$ to $\partial C$, which may be arbitrarily bad. Hence, Garber and Hazan (2015) show that, when the function is strongly-convex, the Frank-Wolfe iterates (with short-step-size at least) converge in $O(1/T^2)$, a convergence rate that in the early iterations can effectively beat badly conditioned linear rates. In Theorem 2.4, we show how $(\alpha, q)$ uniform convexity also allows to continuously interpolate between the $O(1/T)$ and $O(1/T^2)$ rates.

Proof Sketch. We now provide an informal discussion as to why uniformly convexity of $C$ leads to accelerated convergence rates under the classical assumptions that $x^* \in \partial C$ and $||\nabla f(x^*)|| > c > 0$. Formal arguments are developed in the proof of Theorem 2.1. The main insight is the fact that the constraint set around $x^*$ being curved helps to accelerate the convergence of Frank-Wolfe iterates, under suitable step-size choices. Global strong convexity is then just one specific quantification of the curvature of the boundary of a constraint set $C$ that is subsumed by the quantification given by uniform convexity.

The key point is that if $C$ 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 when the constraint set is a polytope. In Figure 1 we show various such behaviors. Applying the $L$-smoothness to Frank-Wolfe iterates, the classical iteration inequality is of the form

\[
 f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \gamma \langle -\nabla f(x_t); v_t - x_t \rangle + \frac{\gamma^2}{2} L ||x_t - v_t||^2. 
\]

The non-negative quantity $\langle -\nabla f(x_t); v_t - x_t \rangle$ participates in guaranteeing the function decrease, counter-balanced with $||x_t - v_t||^2$. Hence when the iterates are close to the optimum, the quantity $||x_t - v_t||$ may be large, but so $\langle -\nabla f(x_t); v_t - x_t \rangle$ will be (see (Garber and Hazan, 2015) for a similar discussion). The convergence rate then depends on specific relative quantification of these various quantities. Strong-convexity of the set is just one quantification of this behavior that can be easily dealt with. Here we present the generalization with uniform convexity.

Figure 1: $v_{\text{strong}}^{FW}$, $v_{\text{uni}}^{FW}$, $v_{\text{poly}}^{FW}$ represents the various Frank-Wolfe vertices from the strongly convex set $C_0$, the uniformly convex set $C_1$ and the polytope $C_2$. 
2.1 Interpolating linear and sublinear rates

Our proof depends on the property of the set at the optimum of (1). In particular, it relies on the characterization of extrema of uniformly convex sets developed by (Dunn, 1979). For a potential solution \( x^* \), we will consider

\[
a_{x^*}(\sigma) \triangleq \inf_{x \in \mathcal{C}} \langle \nabla f(x^*); x - x^* \rangle.
\]

and Dunn (1979) relates some lower bounds on \( a_{x^*}(\sigma) \) with convergence rates of Frank-Wolfe. In particular, when there exists \( \alpha > 0 \) such that \( a_{x^*}(\sigma) \geq \alpha \sigma^2 \) for all \( \sigma \), the Frank-Wolfe algorithm converges linearly, under appropriate line-search rules. This result of (Dunn, 1979) subsumes that of (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970), since for a strongly convex set \( \mathcal{C} \) when \( \| \nabla f(x) \| > c > 0 \), there exists \( \alpha > 0 \) such that for any \( x^* \in \partial \mathcal{C} \) and any \( \sigma > 0 \) we have \( a_{x^*}(\sigma) \geq 2 \alpha \sigma^2 \). Finally (Dunn, 1979, Theorem 3.6.) shows that when \( \mathcal{C} \) is \((\alpha, q)\)-uniformly convex with \( q \geq 2 \), for all \( \sigma > 0 \) we have

\[
a_{x^*}(\sigma) \geq 2 \alpha \| \nabla f(x^*) \| \sigma^q.
\]

However, to our knowledge, no accelerated convergence rate of the Frank-Wolfe algorithm is known under such an assumption, or when the set is uniformly but not strongly convex. We fill this gap in Theorem 2.1 below.

**Theorem 2.1.** Consider a \( \mathcal{L} \)-smooth function \( f \) and \( \mathcal{C} \) an \((\alpha, q)\)-uniformly convex set with \( q > 2 \). Assume \( \| \nabla f(x^*) \| \geq c > 0 \). For simplicity, we also assume that \( h_0 = f(x_0) - f(x^*) \leq 1 \), (which is true up to a simple burn-in phase). Then the iterates of Frank-Wolfe, with short step-size as in Algorithm 1 or exact line search, satisfy

\[
\begin{align*}
    f(x_T) - f^* &\leq \frac{M}{(T + 2)^1/(2(q-1))} \quad \text{when } q > 2, \\
    f(x_T) - f^* &\leq \left(1 - \frac{1}{2 \mathcal{L} \mathcal{H}^2}\right)^T h_0 \quad \text{when } q = 2,
\end{align*}
\]

with

\[
M = \max \left\{ h_0 2^{1/\eta}; c_0 (2 \mathcal{L} \mathcal{H}^2)^{1/\eta} \right\}, \text{ where } \eta = 1 - 2/(q(q-1)) \text{ (which is in } [0, 1] \text{ when } q > 2).
\]

\[
\mathcal{H} = 2 \cdot \max \left\{ \left( \frac{L}{(2\alpha \mathcal{C})} \right)^{1/(q-1)} \left(2\alpha \mathcal{C} \right)^{-1/(q(q-1))}, \left(2\alpha \mathcal{C} \right)^{-1/q} \right\}
\]

and \( c_\eta = \left(5(\exp(1/(3\eta)) - 1)\right)^{1/\eta} \). For the sake of clarity, note that we assume also \( h_0 \leq \left( \mathcal{L} \mathcal{H}^2 \right)^{1-2/(q(q-1))} \), when this is not the case, there is a brief initial linear convergence regime.

**Remark 2.2.** When \( q \) goes to +\( \infty \), \( \eta \) converges to 1 and we recover the classic sublinear convergence rate of \( O(1/T) \) on general compact convex sets.

**Proof.** From (Dunn, 1979, Theorem 3.4), when \( \mathcal{C} \) is \((\alpha, q)\)-uniformly convex, we have for \( x^* \) the solution of (1) and \( \| x - x^* \| \) in particular

\[
a_{x^*}(\| x - x^* \|) \geq 2 \alpha \| \nabla f(x^*) \| \| x - x^* \| ^q \geq 2 \alpha \| x - x^* \| ^q.
\]

Condition (2) is the basis for two important inequalities: one that upper bounds \( \| x - x^* \| \) in terms of \( f(x) - f(x^*) \) and another that upper bounds \( \| v_t - x^* \| \) in terms of \( \| x^* - x_t \| \), where \( v_t \) is the Frank-Wolfe vertex from iterate \( x_t \). These two inequalities rely of convexity, \( \mathcal{L} \)-smoothness and (2), but do not rely on strong convexity of the function \( f \).

By convexity of \( f \) and by definition of \( a_{x^*}(\cdot) \), we have for any \( x \in \mathcal{C} \)

\[
f(x) - f(x^*) \geq \langle \nabla f(x^*); x - x^* \rangle \geq a_{x^*}(\| x - x^* \|) \geq 2 \alpha \| x - x^* \| ^q.
\]
Note that by optimality of the Frank-Wolfe vertex $v_t$, we have $\nabla f(x_t) \leq \nabla f(x^*)$. Hence, combining that with Cauchy-Schwartz and the definition of $a_{x^*}(\cdot)$, we get

$$||\nabla f(x^*) - \nabla f(x_t)|| ||v_t - x^*|| \geq \langle \nabla f(x^*) - \nabla f(x_t); v_t - x^* \rangle + \langle \nabla f(x_t); v_t - x^* \rangle \leq 0$$

Then, $L$-smoothness applied to the left hand side leaves us with

$$||x_t - x^*|| \geq \frac{2c}{L}||v_t - x^*||^{q-1},$$

and a triangular inequality gives

$$||x_t - v_t|| \leq ||v_t - x^*|| + ||x^* - x_t||$$

$$||x_t - v_t|| \leq \left( \frac{L}{2c} \right)^{1/(q-1)} ||x_t - x^*||^{1/(q-1)} + ||x^* - x_t||.$$

Finally applying (3) with $x = x_t$, we have $||x_t - x^*|| \leq \left( \frac{1}{2c} \right)^{1/q} h_t^{1/q}$ which leads to

$$||x_t - v_t|| \leq \left( \frac{L}{2c} \right)^{1/(q-1)} \left( \frac{1}{2c} \right)^{1/(q(q-1))} h_t^{1/(q(q-1))} + \left( \frac{1}{2c} \right)^{1/q} h_t^{1/q}.$$

We can simplify this previous expression, and we assumed without loss of generality (i.e. up to a burning-phase) that $h_t \leq 1$, which implies for $q \geq 2$ that $h_t^{1/(q(q-1))} \geq h_t^{1/q}$.

With $H \triangleq 2 \cdot \max\left\{ \left( \frac{L}{2c} \right)^{1/(q-1)} \left( \frac{1}{2c} \right)^{1/(q(q-1))}, \left( \frac{1}{2c} \right)^{1/q} \right\}$, we then have

$$||x_t - v_t|| \leq H h_t^{1/(q(q-1))}.$$

We now plug this last expression in the classical descent guarantee given by $L$-smoothness

$$h_{t+1} \leq (1 - \gamma) h_t + \gamma^2 \frac{L}{2} ||v_t - x_t||^2$$

$$h_{t+1} \leq (1 - \gamma) h_t + \gamma^2 \frac{L}{2} h_t^{2/(q(q-1))}.$$

The optimal decrease $\gamma \in [0, 1]$ is $\gamma^* = \min\left\{ \frac{h_t^{1-2/(q(q-1))}}{LH^2}, 1 \right\}$. When $\gamma^* = 1$, or equivalently $h_t \geq (LH^2)^{1-2/(q(q-1))}$, we have $h_{t+1} \leq h_t/2$. In other words, for the very first iterations, there is a brief linear convergence regime which we skip to simplify the final bound.

Otherwise, when $\gamma^* \leq 1$, which is true at least for $t \geq \left\lceil \frac{\log(h_0) - (1-2/(q(q-1))) \log(LH^2)}{\log(2)} \right\rceil$, we have

$$h_{t+1} \leq h_t \left( 1 - \frac{1}{2LH^2} h_t^{1-2/(q(q-1))} \right).$$

When $q = 2$, this corresponds to the strongly convex case and we recover the classical linear-convergence regime. Otherwise we conclude using Lemma 2.6 that the rate is $O\left( 1/T^{1/(1-2/(q(q-1)))} \right)$.

**Remark 2.3.** The theorem holds locally and goes beyond the global and local uniformly convex case, in the sense that if for the optimum $x^* \in \partial C$, there exists $A > 0$ and $q > 2$ such that for all $\sigma > 0$

$$a_{x^*}(\sigma) \geq A \sigma^q,$$

then the same convergence rates holds. Indeed in the proof, we access the uniform convexity property only via the lower-bound property on $a_{x^*}(\cdot)$. 

6
2.2 Interpolating Sublinear Rates

When the function is $\mu$-strongly convex, we now prove that uniform convexity of the set also leads to improved sub-linear convergence rates. When $q = 2$, this corresponds to the strong convexity of $C$ for which a known convergence rate is $O(1/T^2)$ (Garber and Hazan, 2015). We show how to continuously interpolate from $O(1/T^2)$ to $O(1/T)$, when varying the parameter of $(\alpha, q)$-uniform convexity in $[2, 3]$.

**Theorem 2.4.** Consider a $L$-smooth and $\mu$-strongly convex function $f$. Assume $C$ is $\gamma$-uniformly strongly convex with respect to $\| \cdot \|$ with $\gamma(\tau) = \alpha q^q$ with $q \in [2, 3]$. Then the iterates of the Frank-Wolfe method, with short step-size as in Algorithm 1 or exact line search, satisfy

$$f(x_t) - f(x^*) \leq M/(T + 2)^{(2/(q-1))},$$

with $M = \max\left\{ h_0 2^{1/\eta}; c_\eta C^{1/\eta} \right\}$, where $\eta = (q-1)/2$ and $C = \frac{\alpha}{4L} \left( \frac{q}{2} \right)^{(3-q)/2}$, where $c_\eta = \left( 5(\exp(1/(3\eta)) - 1) \right)^{1/\eta}$.

**Proof.** We follow (Garber and Hazan, 2015). Recall $h_t = f(x_t) - f(x^*)$ and $x_{t+1} = x_t + \gamma_t(v_t - x_t)$, where $v_t$ is the Frank-Wolfe vertex and $\gamma_t \in [0; 1]$ chosen by exact line-search or short-step size. $L$-smoothness of $f$ then gives that for any $\gamma \in [0, 1]$

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

(5)

By definition of uniform convexity (with $q \geq 2$), we have $\bar{x}_t = \frac{x_t + v_t}{2} - \frac{\alpha}{4}\|v_t - x_t\|^q \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|} \in C$. Then by optimality of $v_t$, we have

$$\langle \nabla f(x_t); v_t - x_t \rangle \leq \langle \nabla f(x_t); \bar{x}_t - x_t \rangle$$

$$\langle \nabla f(x_t); v_t - x_t \rangle \leq -\frac{1}{2}h_t - \frac{\alpha}{4}\|x_t - v_t\|^q \|\nabla f(x_t)\|.$$  

Hence when plugging that into (5), we get

$$h_{t+1} \leq h_t \left( 1 - \frac{\gamma}{2} \right) + \frac{\|v_t - x_t\|^2}{2} \left[ \gamma^2 L - \frac{\alpha}{4}\|\nabla f(x_t)\| \|x_t - v_t\|^{q-2} \right].$$

Note $q = 2$ corresponds to the case where $C$ is strongly convex. If $\gamma = \frac{\alpha}{2L}\|v_t - x_t\|^{q-2}\|\nabla f(x_t)\| \leq 1$ we have

$$h_{t+1} \leq h_t \left( 1 - \frac{\alpha}{4L}\|v_t - x_t\|^{q-2}\|\nabla f(x_t)\| \right).$$

(6)

Note that this choice of $\gamma$ is not optimal but considerably simplifies the analysis. Since the Frank-Wolfe gap is an upper-bound on the primal gap $h_t$, we have

$$h_t \leq \langle -\nabla f(x_t); v_t - x_t \rangle \leq \|\nabla f(x_t)\| \|v_t - x_t\|.$$  

So that (6) becomes

$$h_{t+1} \leq h_t \left( 1 - \frac{\alpha}{4L}h_t^{q-2}\|\nabla f(x_t)\|^{3-q} \right).$$

By strong convexity, we have $\|\nabla f(x)\| \geq \sqrt{\frac{\mu}{2}} \sqrt{h_t}$, hence

$$h_{t+1} \leq h_t \left( 1 - \frac{\alpha}{4L} \left( \frac{\mu}{2} \right)^{(3-q)/2} h_t^{(q-1)/2} \right).$$
Otherwise $\frac{\alpha}{2t}||v_t - x_t||^{q-2}||\nabla f(x_t)|| > 1$ and hence with $\gamma = 1$, we have $h_{t+1} \leq h_t/2$, so that

$$h_{t+1} \leq h_t \cdot \max \left\{ \frac{1}{2} : 1 - \frac{\alpha}{4L} \left( \frac{\mu}{2} \right)^{(3-q)/2} h_t^{(q-1)/2} \right\}.$$ 

Finally the asymptotic convergence rate is $O(T^{-2/(q-1)})$ and Corollary 2.7 gives the exact constants. When $q = 2$, we recover the $O(1/T^2)$ rate of (Garber and Hazan, 2015) and when $q = 3$, we recover the classical sublinear rate of $O(1/T)$. ■

**Remark 2.5.** For $q = 2$, (4) corresponds to the $O(1/T^2)$ of (Garber and Hazan, 2015); for $q = 3$ (4) corresponds to the general sublinear rate of $O(1/T)$. There is probably accelerated rates for $q \geq 3$ but it is algebraically much harder to account for.

### 2.3 Technical Lemma

The proofs of Theorems 2.1 and 2.4 involve finding explicit bounds for sequences $(h_t)$ satisfying recursive inequalities of the form,

$$h_{t+1} \leq h_t(1 - Ch_t^\eta), \quad (7)$$

with $\eta < 1$. The explicit solution with $\eta = 1/2$ is given in (Garber and Hazan, 2015) and corresponds to $h_t = O(1/T^2)$, while for $\eta = 1$ we recover the classical sublinear Frank-Wolfe regime of $O(1/T)$.

For a general $\eta$, we have $O(1/T^{1/\eta})$, which can be guessed by solving the differential equation $\dot{h}(t) = -Ch(t)^{\eta+1}$. The following lemma gives an explicit upper bound on $h_t$ satisfying (7).

**Lemma 2.6.** Assume that for some $0 < \eta \leq 1$, the sequence $(h_t)$ satisfies (7), then

$$h_t \leq \frac{M}{(t + 2)^{1/\eta}},$$

where $M = \max\left\{ h_0 2^{1/\eta}; c_\eta O^{-1/\eta}\right\}$ and $c_\eta = \left( 5(\exp(1/(3\eta)) - 1) \right)^{1/\eta}$.

**Proof.** Note that the choice of $\alpha_\eta = \left(\frac{2}{3}\right)^{1/\eta}$ ensures that $\frac{\alpha_\eta M}{(t+2)^{1/\eta}} \leq \frac{M}{(t+3)^{1/\eta}}$ for $t \geq 1$. Let’s now prove by induction on $t$ that $h_t \leq \frac{M}{(t+2)^{1/\eta}}$. Initialization is true if and only if $M \geq h_0 2^{1/\eta}$. Now, assume by induction that $h_t \leq \frac{M}{(t+2)^{1/\eta}}$ for $t \geq 1$. We distinguish two cases. If $h_t \leq \frac{\alpha_\eta M}{(t+2)^{1/\eta}}$, then by the choice of $\alpha_\eta$ we have that

$$h_{t+1} \leq h_t \leq \frac{M}{(T + 3)^{1/\eta}},$$

and the induction if proven. When $h_t \geq \frac{\alpha_\eta M}{(t+2)^{1/\eta}}$ on the other hand, we have a lower bound on $h_t$ that we inject in (7) to get

$$h_{t+1} \leq \frac{M}{(t + 2)^{1/\eta}} \left( 1 - CM^\eta \frac{1}{t + 2} \right)$$

and

$$h_{t+1} \leq \frac{M}{(t + 3)^{1/\eta}} \exp \left( \frac{1}{\eta(t + 2)} \right) \left( 1 - 2CM^\eta \frac{1}{3(t + 2)} \right).$$
In particular for $t \geq 1$, $\frac{1}{\eta(t+2)} \in [0, \frac{1}{3\eta}]$ and by convexity of the exponential we have for $x \in [0, \frac{1}{3\eta}]$, $\exp(x) \leq 3\eta(\exp(1/(3\eta)) - 1)x + 1$ so that
\[
h_{t+1} \leq \frac{M}{(t + k + 1)^{1/\eta}} \left(1 + \frac{3(\exp(1/(3\eta)) - 1)}{t + 2}\right) \left(1 - 2CM^{\eta} \frac{1}{3(t + 2)}\right).
\]

Finally to ensure that for all $t \geq 1$
\[
\left(1 + \frac{3(\exp(1/(3\eta)) - 1)}{t + 2}\right) \left(1 - 2CM^{\eta} \frac{1}{3(t + 2)}\right) \leq 1,
\]
it is sufficient to have $M$ such that
\[
3(\exp(1/(3\eta)) - 1) \leq 2CM^{\eta}/3,
\]
hence $M \geq \left(\frac{5}{C}\right)^{1/\eta} \left(\exp(1/(3\eta)) - 1\right)^{1/\eta}$, which proves the induction in the second case.

**Corollary 2.7.** When $(h_t)$ satisfies for $\eta \in [1/2, 1]$
\[
h_{t+1} \leq h_t \cdot \max\left\{\frac{1}{2}, 1 - Ch_t^{\eta}\right\},
\]
we also have
\[
h_t \leq \frac{M}{(t + 2)^{1/\eta}},
\]
with the same value of $M$ as in Lemma 2.6.

**Proof.** When the maximum is realized on the second argument, the rational is the same as in Lemma 2.6. Otherwise, in the previous induction argument, we would have
\[
h_{t+1} \leq \frac{M}{2(t + 2)^{1/\eta}} \leq \frac{M}{(t + 3)^{1/\eta}},
\]
because $\eta \geq \frac{1}{2}$ and $t \geq 1$.

### 3 Examples of Uniformly Convex Objects

Uniform convexity assumption refines convexity properties of several mathematical objects, such as normed spaces, functions and sets. To analyze the Frank-Wolfe algorithm, we relied on the uniform convexity of sets. In this section, we connect the various uniform convexity notions. It helps to characterize the uniform convexity of classical constraint sets. In Section 3.1, we recall that norm balls of uniformly convex spaces are uniformly convex sets. In particular, we deduce the set uniform convexity of classic norm balls in Section 3.2. In Section 3.3 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).
3.1 Uniformly Convex Spaces

For norm balls, uniform convexity of sets as stated in Definition 1.1 is closely related to uniform convexity of normed spaces (Polyak, 1966; Balashov and Repovs, 2011; Lindenstrauss and Tzafriri, 2013; Weber and Reisig, 2013). Some works establish sharp uniform convexity results for classical normed spaces such as $l_p$, $L_p$ or $C_p$. The practical examples of uniformly convex set that we give stem from classical norm balls, and hence are tightly linked with uniformly convex spaces. However, it is easy to construct uniformly convex sets that are not norm balls. Consider for instance $C = \{ x \mid \|x\|_2 \leq 1 \text{ when } x^{(1)} \leq 0\text{ and } \|x\|_3 \leq 1 \text{ otherwise}\}.$ (Clarkson, 1936; Boas Jr, 1940) define a uniformly convex normed space $(X, \|\cdot\|)$ 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 $X$ 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, see Definition 3.1.

**Definition 3.1 (Modulus of convexity).** The modulus of convexity of the space $(X, \|\cdot\|)$ is defined as

$$\delta_X(\epsilon) = \inf \left\{ 1 - \frac{\|x + y\|}{2} \mid \|x\| \leq 1, \|y\| \leq 1, \|x - y\| \geq \epsilon \right\}.$$

Note that some works consider $2\epsilon$ instead of $\epsilon$ in Definition 3.1. More specifically, a normed-space $X$ is said to be $r$-uniformly convex in the case $\delta_X(\epsilon) \geq C\epsilon^r$. 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. Some sharp results exists for $L_p$ and $l_p$ 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 et al., 1979; Ball et al., 1994). The following gives a straightforward link between the set $\delta_C$ and space $\delta_X$ modulus of convexity.

**Lemma 3.2.** If a normed space $(X, \|\cdot\|)$ is uniformly convex with modulus of convexity $\delta_X(\cdot)$, then its unit norm ball is $4\delta_X(\cdot)$ uniformly convex with respect to $\|\cdot\|$.

**Proof.** Assume $(X, \|\cdot\|)$ is uniformly convex with modulus of convexity $\delta(\cdot)$. Then for any $(x, y, z) \in B_{\|\cdot\|}(1)$, we have by definition $1 - \frac{\|x + y\|}{2} \geq \delta(\|x - y\|)$ and then

$$\frac{x + y}{2} + \delta(\|x - y\|)z \leq \frac{x + y}{2} + \delta(\|x - y\|) \leq 1.$$

3.2 Uniform convexity of some classic norm balls

We now gather the uniform convexity results for $l_p$ and $p$-Schatten unit balls. Note that if the unit ball $B_{\|\cdot\|}(1)$ is $(\alpha, q)$-uniformly convex, then $B_{\|\cdot\|}(r)$ is $(\alpha/r^{q-1}, q)$-uniformly convex.

**Case of $\ell_p$-balls.** When $p \in [1, 2]$, $\ell_p$-ball are strongly convex set and $((p - 1)/2, 2)$-uniformly convex with respect to $\|\cdot\|_p$, see for instance (Hanner et al., 1956, Theorem 2) (Garber and Hazan, 2015, Lemma 4). This corresponds to the results of (Hanner et al., 1956, Theorem 2). 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 where $p, s > 1$. 


Schatten Balls. (Dixmier, 1953; Tomczak-Jaegermann, 1974; Simon et al., 1979; Ball et al., 1994) focus of the uniform convexity of the \((C_p, ||\cdot||_{S(p)})\) spaces which unit ball correspond to 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).

3.3 Uniformly Convex Functions

Uniform convexity is also a property of a convex function and defined as follows.

**Definition 3.3.** A differentiable function \(f\) is \((\mu, r)\)-uniformly convex on a convex set \(C\) if there exists \(r \geq 2\) and \(\mu > 0\) such that for all \((x, y) \in C\)

\[
  f(y) \geq f(x) + \langle \nabla f(x); y - x \rangle + \frac{\mu}{2} ||x - y||^2_r.
\]

We now state the equivalent of (Journée et al., 2010, Theorem 12) for the level sets of uniformly convex functions. This was already used in (Garber and Hazan, 2015) in the case of strongly-convex sets. All the proofs can be found in Appendix A.

**Lemma 3.4.** Let \(f : \mathbb{R}^d \to \mathbb{R}^+\) be a non-negative, \(L\)-smooth and \((\mu, r)\)-uniformly convex function on \(\mathbb{R}^d\), with \(r \geq 2\). Then for any \(w > 0\), the set

\[
  \mathcal{L}_w = \left\{ x \mid f(x) \leq w \right\},
\]

is \((\alpha, r)\)-uniformly convex with \(\alpha = \frac{\mu}{\sqrt{2wL}}\).

Lemma 3.4 restrictively requires smoothness of the uniformly convex function \(f\). Hence we provide the analogous of (Garber and Hazan, 2015, Lemma 3).

**Lemma 3.5.** Consider a finite dimensional normed vector space \((\mathbb{X}, ||\cdot||)\). Assume \(f(x) = ||x||^2\) is \((\mu, s)\)-uniformly convex function (with \(r \geq 2\)) with respect to ||\cdot||. Then the norm balls \(B_{||\cdot||}(r) = \left\{ x \in \mathbb{X} \mid ||x|| \leq r \right\}\) are \((\frac{\mu}{2r}, s)\)-uniformly convex.

These previous lemma hence allow to translate functional uniformly convex results into results for classic balls norms. For instance, (Shalev-Shwartz, 2007, Lemma 17) showed that for \(p \in [1, 2]\) \(f(x) = 1/2 ||x||^2_p\) was \((p - 1)\)-uniformly convex with respect to ||\cdot||_p.

3.4 Local Uniform Convexity

Uniform convexity has also been extensively used in optimal control (Pliś, 1975; Dunn, 1979; Ivanov, 1997; Balashov and Repovs, 2009), with an emphasis on local characterisation of strong convexity (Balashov and Repovs, 2011; Weber and Reichs, 2013). For the sake of simplicity, Theorems 2.1 and 2.4 are stated in terms of a global uniform convexity constant on the domain \(C\). However, the proof of Theorem 2.1 relies on a local uniform convexity assumption (Dunn, 1979). In particular, local characterizations of classic ball constraints would result in different convergence behaviors according to the position of \(x^* \in \partial C\) and would refine our convergence results.
4 Numerical Examples

Uniform convexity is a global assumption. Hence, in Theorem 2.1, we obtain sublinear convergence that do not depend on the specific location of the solution $x^* \in \partial C$. However, some regions of $C$ might be relatively more curved than others and hence exhibit faster convergence rates. This effect can be quantified by Theorem 2.1 provided some local characterisations of the uniformly convex sets.

In Figure 2, in the case of the $\ell_p$ balls with $p > 2$, we vary the approximate location of the optimum $x^*$ in the boundary of the $\ell_p$ balls.

Subfigures 2a-2c are associated to an optimization problem where we placed the solution $x^*$ to (1) 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). This is typically a region of the boundaries of the $\ell_p$ we expect to be curved.

Subfigures 2d-2f, corresponds to the same optimization problem where we place the optimum $x^*$ near the intersection between the half-line generated by $e_1$ and the boundary of the $\ell_p$ balls. This is typically a flat region 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 (subfigures 2b and 2e). However, when the optimum is near the flat location, we indeed observe sublinear convergence rates (e and f). It still becomes linear for $p = 2.1$ with exact line-search, subfigure 2f.

Also, Theorem 2.1 gives accelerated rates when using the Frank-Wolfe algorithm with exact line-search or short-step size. In subfigures 2a and 2d of Figure 2 we show example of Frank-Wolfe convergence when using deterministic line-search. The rates are indeed sublinear in $O(1/T)$. In other words, deterministic line-search generally do not lead to accelerated convergence rates when the sets are uniformly convex.
Figure 2: Solving (1) with Frank-Wolfe algorithm where $f$ is a quadratic with conditioning 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 vary the position of the solution to (1) with respect to the boundaries of the constraint sets. On the first row, we choose the constrained optimum close to the intersection of the set boundary and the line generated by $\sum_i e_i$ (where the $e_i$ form the canonical basis), where $\ell_p$ balls are typically more curved. On the second row, we choose the constrained optimum near the intersection between the set boundary and the line generated by $e_1$, a region where the $\ell_p$ balls are flat. On a line, each plot exhibits the behavior of the Frank-Wolfe algorithm iterates with different step-size strategy: determinist line-search (i.e. $1/(k+1)$), short step sizes and exact line-search. To avoid the oscillating behavior of Frank-Wolfe gap, the $y$-axis represents $\min_{k=1,\ldots,T} g(x_k)$ where $g(\cdot)$ is the Frank-Wolfe gap and $T$ the number of iterations.

5 Conclusion

Our results fill the gap between known convergence rates for the Frank Wolfe algorithm. Qualitatively, they also mean that it is the curvature of the constraint set that accelerates the convergence of the Frank-Wolfe algorithm, not just strong-convexity. This emphasis on curvature echoes works in other settings (Huang et al., 2016). For the sake of theory, the results could be immediately refined by measuring the local curvature of convex bodies with more sophisticated tools than uniform convexity (Schneider, 2015).

From a more practical perspective, uniform convexity encompasses ubiquitous structures of constraint sets appearing in machine learning and signal processing. In applications where the (e.g. regularization) constraints are likely to be active, the assumption that $|\nabla f(x^*)| \geq c > 0$ is not restrictive and the value of $c$ quantifies the relevance of the constraints.
References

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

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

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

M. V. Balashov and D. Repovs. Uniform convexity and the splitting problem for selections. Journal of mathematical analysis and applications, 360(1):307–316, 2009.

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

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

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

E. Blanc. Les ensembles surconvexes plans. In Annales scientifiques de l’École Normale Supérieure, volume 60, pages 215–246, 1943.

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

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

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

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

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

K. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. ACM Transactions on Algorithms (TALG), 6(4):65, 2010.

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

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

J. Dixmier. Formes linéaires sur un anneau d’opérateurs. Bulletin de la Société Mathématique de France, 81:9–39, 1953.
M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Artificial Intelligence and Statistics*, pages 327–336, 2012.

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

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

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

F. Futami, Z. Cui, I. Sato, and M. Sugiyama. Bayesian posterior approximation via greedy particle optimization, 2018.

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

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

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

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

J. Guélat and P. Marcotte. Some comments on wolfe’s “away step”. *Mathematical Programming*, 35(1):110–119, 1986.

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

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

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

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

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

R. Huang, T. Lattimore, A. György, and C. Szepesvári. 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, 2016.

G. E. Ivanov. Saddle point for differential games with strongly convex-concave integrand. *Mathematical Notes*, 62(5):607–622, 1997.

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

M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, number CONF, pages 427–435, 2013.
M. Jaggi and M. Sulovský. A simple algorithm for nuclear norm regularized problems. 2010.

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

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

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank–Wolfe optimization variants. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 496–504. Curran Associates, Inc., 2015.

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

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

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

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

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

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

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

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

F. Pedregosa, A. Askari, G. Negiar, and M. Jaggi. Step-size adaptivity in projection-free optimization. *arXiv preprint arXiv:1806.05123*, 2018.

J. Pena and D. Rodriguez. Polytope conditioning and linear convergence of the frank–wolfe algorithm. *Mathematics of Operations Research*, 2018.

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

A. Pliś. Uniqueness of optimal trajectories for non-linear control systems. In *Annales Polonici Mathematici*, volume 4, pages 397–401, 1975.

B. T. Polyak. 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, 1966.

R. Schneider. Curvatures of typical convex bodies—the complete picture. *Proceedings of the American Mathematical Society*, 143(1):387–393, 2015.
G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev. Instance-level video segmentation from object tracks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3678–3687, 2016.

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

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

B. Simon et al. Trace ideals and their applications. 1979.

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

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

T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*, 2018.

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

Proof of Lemma 3.4. The proof exactly that of (Journé et al., 2010, Theorem 12), replacing \(||x - y||^r\) with \(||x - y||^r\) but we state it for the sake of completeness. Consider \(w_0 > 0, (x, y) \in \mathcal{L}_w\) and \(\gamma \in [0, 1]\). We denote \(z = \gamma x + (1 - \gamma)y\). For \(u \in \mathbb{R}^d\), by \(L\)-smoothness applied at \(z\) and at \(x^*\) (the unconstrained optimum of \(f\)), we have

\[
    f(z + u) \leq f(z) + (\nabla f(z), u) + \frac{L}{2}||u||^2
\]

\[
    \leq f(z) + ||\nabla f(z)|| \cdot ||u|| + \frac{L}{2}||u||^2
\]

\[
    \leq f(z) + \sqrt{2Lf(z)}||u|| + \frac{L}{2}||u||^2 = \left(\sqrt{f(z)} + \sqrt{\frac{L}{2}||u||}\right)^2.
\]

Note that uniform convexity of \(f\) implies that

\[
    f(z) \leq \gamma f(x) + (1 - \gamma)f(y) - \frac{\mu}{2}\gamma(1 - \gamma)||x - y||^r
\]

In particular then, because \(x, y \in \mathcal{L}_w\), we have \(f(z) \leq w - \frac{\mu}{2}\gamma(1 - \gamma)||x - y||^r\) so that

\[
    f(z + u) \leq \left(\sqrt{w - \frac{\mu}{2}\gamma(1 - \gamma)||x - y||^r} + \sqrt{\frac{L}{2}||u||}\right)^2
\]

Leveraging on the concavity of the square-root, we get

\[
    f(z + u) \leq \left(\sqrt{w - \frac{\mu}{4\sqrt{w}}\gamma(1 - \gamma)||x - y||^r} + \sqrt{\frac{L}{2}||u||}\right)^2.
\]

Hence for any \(u\) such that

\[
    ||u|| = \frac{\mu}{2\sqrt{2wL}}\gamma(1 - \gamma)||x - y||^r,
\]

we have \(z + u \in \mathcal{L}_w\). Hence \(\mathcal{L}_w\) is a \((\frac{\mu}{2\sqrt{2wL}}, r)\)-uniformly convex set. 

Proof of Lemma 3.5. The proof follows exactly that of (Garber and Hazan, 2015, Lemma 3) which itself follows that of (Journé et al., 2010, Theorem 12), where operations involving \(L\)-smoothness are replaced by an application of the triangular inequality.

Let’s consider \(s \geq 2, (x, y) \in B_{\|\cdot\|}(r)\) and \(\gamma \in [0, 1]\). We denote \(z = \gamma x + (1 - \gamma)y\). For \(u \in \mathbb{X}\), applying successively triangular inequality and \((\mu, s)\)-uniform convexity of \(f(x) = \|x\|^2\), we get

\[
    f(z + u) = \|z + u\|^2 \leq \left(\sqrt{f(z)} + ||u||\right)^2
\]

\[
    \leq \left(\sqrt{r^2 - \frac{\mu}{2}\gamma(1 - \gamma)||x - y||^s} + ||u||\right)^2.
\]

We then use concavity of the square root as before to get

\[
    \|z + u\|^2 \leq \left(r - \frac{\mu}{4r}\gamma(1 - \gamma)||x - y||^s + ||u||\right)^2.
\]

In particular, for \(u \in \mathbb{X}\) such that \(||u|| = \frac{\mu}{4r}\gamma(1 - \gamma)||x - y||^s\), we have \(z + u \in B_{\|\cdot\|}(r)\). Hence \(B_{\|\cdot\|}(r)\) is \((\frac{\mu}{4r}, s)\)-uniformly convex with respect to \(||\cdot||\).