Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets

Dan Garber DANGAR@TX.TECHNION.AC.IL
Technion - Israel Institute of Technology

Elad Hazan EHAZAN@CS.PRINCETON.EDU
Princeton University

Abstract

The Frank-Wolfe method (a.k.a. conditional gradient algorithm) for smooth optimization has regained much interest in recent years in the context of large scale optimization and machine learning. A key advantage of the method is that it avoids projections - the computational bottleneck in many applications - replacing it by a linear optimization step. Despite this advantage, the known convergence rates of the FW method fall behind standard first order methods for most settings of interest. It is an active line of research to derive faster linear optimization-based algorithms for various settings of convex optimization.

In this paper we consider the special case of optimization over strongly convex sets, for which we prove that the vanila FW method converges at a rate of $\frac{1}{t^2}$. This gives a quadratic improvement in convergence rate compared to the general case, in which convergence is of the order $\frac{1}{t}$, and known to be tight. We show that various balls induced by ℓ_p norms, Schatten norms and group norms are strongly convex on one hand and on the other hand, linear optimization over these sets is straightforward and admits a closed-form solution. We further show how several previous fast-rate results for the FW method follow easily from our analysis.

1. Introduction

The Frank-Wolfe method, originally introduced by Frank and Wolfe in the 1950's (Frank & Wolfe, 1956), is a *first order* method for the minimization of a smooth convex function over a convex set. Its main advantage in large-scale problems is that it is a first-order and projection-free method - i.e. the algorithm proceeds by iteratively solving a linear optimization problem and remaining inside the feasible domain. For matrix completion problems, metric learning, sparse PCA, structural SVM and other large-scale machine learning problems, this feature enabled the derivation of algorithms that are practical on one hand and come with provable convergence rates on the other (Jaggi & Sulovský, 2010; Lacoste-Julien et al., 2013; Dudík et al., 2012; Harchaoui et al., 2012; Hazan & Kale, 2012; Shalev-Shwartz et al., 2011; Laue, 2012).

Despite its empirical success, the main drawback of the method is its relatively slow convergence rate in comparison to optimal first order methods. The convergence rate of the method is on the order of $1/t$ where t is the number of iterations, and this is known to be tight. In contrast, Nesterov's accelerated gradient descent method gives a rate of $1/t^2$ for general convex smooth problems and a rate $e^{-\Theta(t)}$ is known for smooth and strongly convex problems. The following question arises: are there projection-free methods with convergence rates matching that of projected gradient-descent and its extensions?

Motivated by this question, in this work we advance the line of research for faster convergence rates of projection free methods. We prove that in case both the objective function and the feasible set are strongly convex (in fact a slightly weaker assumption than strong convexity of the objective is required), the vanilla Frank-Wolfe method converges at an accelerated rate of $1/t^2$. The improved convergence rate is independent of the dimension. This is also the first convergence result for the FW that we are aware of that achieves a rate that is between the standard $1/t$ rate and a linear rate. We further show how the analysis used to prove the latter result enables to easily derive previous fast convergence rates for the FW method.

We motivate the study of optimization over strongly convex sets by demonstrating that various norms that serve as popular regularizes in machine learning problems, including ℓ_p norms, matrix Schatten norms and matrix group norms,

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

give rise to strongly convex sets. We further show that indeed linear optimization over these sets is straightforward to implement and admits a closed-form solution. Hence the FW method is appealing for solving optimization problems with such constraints, such as regularized linear regression.

1.1. Related Work

The Frank-Wolfe method dates back to the original work of Frank and Wolfe (Frank & Wolfe, 1956) which presented an algorithm for minimizing a quadratic function over a polytope using only linear optimization steps over the feasible set. Recent results by Clarkson (Clarkson, 2008), Hazan (Hazan, 2008) and Jaggi (Jaggi, 2013) extend the method to smooth convex optimization over the *simplex*, *spectrahedron* and arbitrary convex and compact sets respectively.

It was shown in numerous works that the convergence rate of the method is on the order of $1/t$ and that it could not be improved in general, even if the objective function is strongly convex for instance, as shown in (Clarkson, 2008; Hazan, 2008; Jaggi, 2013), even though it is known that in this case, the projected gradient method achieves an exponentially fast convergence rate.

Over the past years, several results tried to improve the convergence rate of the Frank-Wolfe method under various assumptions. GuéLat and Marcotte (GuéLat & Marcotte, 1986) showed that in case the objective function is strongly convex and the feasible set is a polytope, then in case the optimal solution is located in the interior of the set, the FW method converges exponentially fast. A similar result was presented in the work of Beck and Teboulle (Beck & Teboulle, 2004) who considered a specific problem they refer to a *the convex feasibility problem* over an arbitrary convex set. They also obtained a linear convergence rate under the assumption that an optimal solution that is far enough from the boundary of the set exists.

Recently, Garber and Hazan (Garber & Hazan, 2013a) gave the first natural linearly-converging FW variant without any restricting assumptions on the location of the optimum. They showed that a variant of the Frank Wolfe method with the addition of *away steps* converges exponentially fast in case the objective function is strongly convex and the feasible set is a polytope. In follow-up work, Jaggi and Lacoste-Julien (Lacoste-Julien & Jaggi, 2013) gave a refined analysis of an algorithm presented in (GuéLat & Marcotte, 1986) which also uses away steps and showed that it also converges exponentially fast in the same setting as the Garber-Hazan result. Also relevant in this context is the work of Ahipasaoglu, Sun and Todd (Ahipasaoglu et al., 2008) who showed that in the specific case of minimizing a smooth and strongly convex function over the unit simplex, a variant of the Frank-Wolfe method that also uses away steps converges with a linear rate.

In a different line of work, Migdalas and recently Lan (Migdalas, 1994; Lan, 2013) considered the Frank-Wolfe algorithm with a stronger optimization oracle that is able to solve quadratic problems over the feasible domain. They show that in case the objective function is strongly convex then exponentially fast convergence is attainable. However, in most settings of interest, an implementation of such a non-linear oracle is computationally much more expensive than the linear oracle, and the key benefit of the Frank-Wolfe method is lost.

In the specific case that the feasible set is strongly convex, an assumption also made in this paper, Levitin and Polyak showed in their classical work (Levitin & Polyak, 1966) that under the restrictive assumption that the norm of the gradient of the objective function is lower bounded by a constant everywhere in the feasible set, the FW method converges with an exponential rate. The same result appeared in following works by Demyanov and Rubinov (Demyanov & Rubinov, 1970) and Dunn (Dunn, 1979), both also requiring that the magnitude of the gradients is lower bounded by a constant everywhere in the feasible set. As we later show, the lower bound requirement on the gradients is in a sense much stronger than requiring that the objective function is strongly convex. Under our assumption however, which is slightly weaker than strong convexity of the objective, the gradient may become arbitrarily small on the feasible set.

We summarize previous convergence rate results for the standard FW method in Table 1.1.

2. Preliminaries

2.1. Smoothness and Strong Convexity

For the following definitions let $\mathbf{E}$ be a finite vector space and $\|\cdot\|, \|\cdot\|_*$ be a pair of dual norms over $\mathbf{E}$.

Definition 1 (smooth function)**.** *We say that a function $f : \boldsymbol{E} \to \mathbb{R}$ is β smooth over a convex set $\mathcal{K} \subset \boldsymbol{E}$ with respect to $\|\cdot\|$ if for all $x, y \in \mathcal{K}$ it holds that*

$$f(y) \leq f(x) + \nabla f(x) \cdot (y - x) + \frac{\beta}{2}\|x - y\|^2.$$

Definition 2 (strongly convex function)**.** *We say that a function $f : \boldsymbol{E} \to \mathbb{R}$ is α-strongly convex over a convex set $\mathcal{K} \subset \boldsymbol{E}$ with respect to $\|\cdot\|$ if it satisfies the following two equivalent conditions*

1. $\forall x, y \in \mathcal{K}$:

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}\|x - y\|^2.$$

Reference	Feasible set $\mathcal{K}$	Objective function f	Location of x^*	Conv. rate
(Jaggi, 2013)	convex	convex	unrestricted	$1/t$
(GuéLat & Marcotte, 1986)	polytope	strongly convex	interior	$\exp(-\Theta(t))$
(Beck & Teboulle, 2004)	convex	$f(x) = \|Ax - b\|_2^2$	interior	$\exp(-\Theta(t))$
(Levitin & Polyak, 1966) (Demyanov & Rubinov, 1970) (Dunn, 1979)	strongly convex	$\|\nabla f(x)\| \geq c > 0 \quad \forall x \in \mathcal{K}$	unrestricted	$\exp(-\Theta(t))$
this paper	strongly convex	strongly convex	unrestricted	$1/t^2$

Table 1. Comparison of convergence rates for the Frank-Wolfe method under different assumptions. We denote the optimal solution by x^*. Note that since all results assume smoothness of the function we omit it from column 3.

2. $\forall x, y \in \mathcal{K}, \gamma \in [0,1]$:

$$\begin{aligned} f(\gamma x + (1-\gamma)y) &\leq \gamma f(x) + (1-\gamma) f(y) \\ &- \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|^2. \end{aligned}$$

The above definition (part 1) combined with first order optimality conditions imply that for a α-strongly convex function f, if $x^* = \arg\min_{x \in \mathcal{K}} f(x)$, then for any $x \in \mathcal{K}$

$$f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2. \tag{1}$$

Eq. (1) further implies that the magnitude of the gradient of f at point x, $\|\nabla f(x)\|_*$ is at least of the order of the square-root of the approximation error at x, $f(x) - f(x^*)$. This follows since

$$\begin{aligned} \sqrt{\frac{2}{\alpha}\left(f(x) - f(x^*)\right)} \cdot \|\nabla f(x)\|_* &\geq \|x - x^*\| \cdot \|\nabla f(x)\|_* \\ &\geq (x - x^*) \cdot \nabla f(x) \\ &\geq f(x) - f(x^*), \end{aligned}$$

where the first inequality follows from (1), the second from Holder's inequality and the third from convexity of f. Thus we have that at any point $x \in \mathcal{K}$ it holds that

$$\|\nabla f(x)\|_* \geq \sqrt{\frac{\alpha}{2}} \cdot \sqrt{f(x) - f(x^*)}. \tag{2}$$

We will show that this property, that is in fact weaker than strong convexity, combined with an additional property of the convex set that we define next, allows to obtain the faster rates [1].

Definition 3 (strongly convex set)**.** *We say that a convex set $\mathcal{K} \subset \boldsymbol{E}$ is α-strongly convex with respect to $\|\cdot\|$ if for any $x, y \in \mathcal{K}$, any $\gamma \in [0,1]$ and any vector $z \in \boldsymbol{E}$ such that $\|z\| = 1$, it holds that*

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

That is, $\mathcal{K}$ contains a ball of of radius $\gamma(1-\gamma)\frac{\alpha}{2}\|x-y\|^2$ induced by the norm $\|\cdot\|$ centered at $\gamma x + (1-\gamma)y$.

[1] In this work we assume that the convex set $\mathcal{K}$ is full-dimensional. In case this assumption does not hold, e.g. if the convex set is the unit simplex, then Eq. (2) holds even if we replace $\nabla f(x)$ with $P_{S(\mathcal{K})}[\nabla f(x)]$ where $P_{S(\mathcal{K})}$ denotes the projection operator onto the smallest subspace that contains $\mathcal{K}$.

2.2. The Frank-Wolfe Algorithm

The Frank-Wolfe algorithm, also known as the *conditional gradient algorithm*, is an algorithm for the minimization of a convex function $f : \mathbf{E} \to \mathbb{R}$ which is assumed to be β_f-smooth with respect to a norm $\|\cdot\|$, over a convex and compact set $\mathcal{K} \subset \mathbf{E}$. The algorithm implicitly assumes that the convex set $\mathcal{K}$ is given in terms of a linear optimization oracle $\mathcal{O}_\mathcal{K} : \mathbf{E} \to \mathcal{K}$ which given a linear objective $c \in \mathbf{E}$ returns a point $x = \mathcal{O}_\mathcal{K}(c) \in \mathcal{K}$ such that $x \in \arg\min_{y \in \mathcal{K}} y \cdot c$. The algorithm is given below. The algorithm proceeds in iterations, taking on each iteration t the new iterate x_{t+1} to be a convex combination between the previous feasible iterate x_t and a feasible point that minimizes the dot product with the gradient direction at x_t, which is generated by invoking the oracle $\mathcal{O}_\mathcal{K}$ with the input vector $\nabla f(x_t)$. There are various ways to set the parameter that controls the convex combination η_t in order to guarantee convergence of the method. The option that we choose here is the optimization of η_t via a simple line search rule, which is straightforward and computationally cheap to implement.

Algorithm 1 Frank-Wolfe Algorithm

1: Let x_0 be an arbitrary point in $\mathcal{K}$.
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad p_t \leftarrow \mathcal{O}_\mathcal{K}(\nabla f(x_t))$.
4: $\quad \eta_t \leftarrow \arg\min_{\eta \in [0,1]} \eta(p_t - x_t) \cdot \nabla f(x_t) + \eta^2 \frac{\beta_f}{2}\|p_t - x_t\|^2$.
5: $\quad x_{t+1} \leftarrow x_t + \eta_t(p_t - x_t)$.
6: **end for**

The following theorem states the well-known convergence rate of the Frank-Wolfe algorithm for smooth convex minimization over a compact and convex set, without any further assumptions. A proof is given in the appendix for completeness though similar proofs could also be found in (Levitin & Polyak, 1966; Jaggi, 2013).

Theorem 1. *Let $x^* \in \arg\min_{x\in\mathcal{K}} f(x)$ and denote $D_{\mathcal{K}} = \max_{x,y\in\mathcal{K}} \|x-y\|$ (the diameter of the set with respect to $\|\cdot\|$). For every $t \geq 1$ the iterate x_t of Algorithm 1 satisfies*

$$f(x_t) - f(x^*) \leq \frac{8\beta_f D_{\mathcal{K}}^2}{t} = O\left(\frac{1}{t}\right).$$

2.3. Our Results

In this work, we consider the case in which the function to optimize f is not only β_f-smooth with respect to $\|\cdot\|$ but also α_f-strongly convex with respect to $\|\cdot\|$ (we relax this assumption a bit in subsection 4.3). We further assume that the feasible set $\mathcal{K}$ is $\alpha_{\mathcal{K}}$-strongly convex with respect to $\|\cdot\|$. Under these two additional assumptions alone we prove the following theorem.

Theorem 2. *Let $x^* = \arg\min_{x\in\mathcal{K}} f(x)$ and let $M = \frac{\sqrt{\alpha_f}\alpha_{\mathcal{K}}}{8\sqrt{2}\beta_f}$. Denote $D_{\mathcal{K}} = \max_{x,y\in\mathcal{K}} \|x-y\|$. For every $t \geq 1$ the iterate x_t of Algorithm 1 satisfies*

$$f(x_t) - f(x^*) \leq \frac{\max\{\frac{9}{2}\beta_f D_{\mathcal{K}}^2, 18M^{-2}\}}{(t+2)^2} = O\left(\frac{1}{t^2}\right).$$

3. Proof of Theorem 2

We denote the approximation error of the iterate x_t produced by the algorithm by h_t. That is $h_t = f(x_t) - f(x^*)$ where $x^* = \arg\min_{x\in\mathcal{K}} f(x)$.

To better illustrate our results, we first shortly revisit the proof technique of Theorem 1. The main observation to be made is the following:

$$\begin{aligned}
&h_{t+1} = f(x_t + \eta_t(p_t - x_t)) - f(x^*) \leq \\
&h_t + \eta_t(p_t - x_t)\cdot\nabla f(x_t) + \frac{\eta_t^2\beta_f}{2}\|p_t - x_t\|^2 \leq \\
&h_t + \eta_t(x^* - x_t)\cdot\nabla f(x_t) + \frac{\eta_t^2\beta_f}{2}\|p_t - x_t\|^2 \leq \\
&(1-\eta_t)h_t + \frac{\eta_t^2\beta_f}{2}\|p_t - x_t\|^2, \qquad (3)
\end{aligned}$$

where the the first inequality follows from the smoothness of f, the second from the optimality of p_t and the third from convexity of f. Choosing η_t to be roughly $1/t$ yields the convergence rate of $1/t$ stated in Theorem 1. This rate cannot be improved in general since while the so-called *duality gap* $(x_t - p_t)\cdot\nabla f(x_t)$ could be arbitrarily small (as small as $(x_t - x^*)\cdot\nabla f(x_t)$), the quantity $\|p_t - x_t\|$ may remain as large as the diameter of the set. Note that in case f is strongly-convex, then according to Eq. (1) it holds that x_t converges to x^* and thus according to Eq. (3) it suffices to solve the inner linear optimization problem in Algorithm 1 on the intersection of $\mathcal{K}$ and a small ball centered at x_t. As a result the quantity $\|p_t - x_t\|^2$ will be proportional to the approximation error at time t, and a linear convergence rate will be attained. However in general, under the linear oracle assumption, we have no way to solve the linear optimization problem over the intersection of $\mathcal{K}$ and a ball without greatly increasing the number of calls to the linear oracle, which is the most expensive step in many settings.

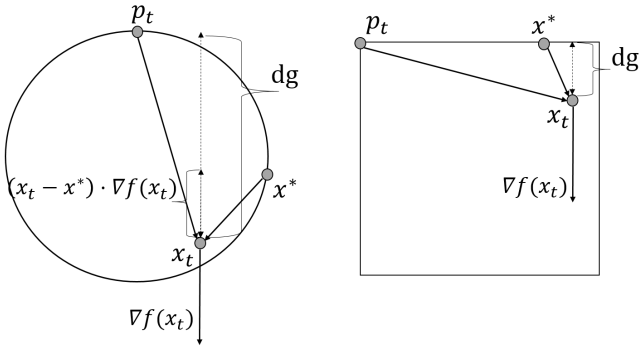

Figure 1. For strongly convex sets, as in the left picture, the duality gap (denoted dg) increases with $\|p_t - x_t\|^2$, which accelerates the convergence of the Frank-Wolfe method. As shown in the picture on the right, this property clearly does not hold for arbitrary convex sets.

In case the feasible set $\mathcal{K}$ is strongly convex, then the main observation to be made is that while the quantity $\|p_t - x_t\|$ may still be much larger than $\|x^* - x_t\|$ (the distance to the optimum), in this case, the *duality gap* must also be large, which results in faster convergence. This observation is illustrated in Figure 1 and given formally in Lemma 1.

Lemma 1. *On any iteration t of Algorithm 1 it holds that*

$$h_{t+1} \leq h_t \cdot \max\{\frac{1}{2}, 1 - \frac{\alpha_K\|\nabla f(x_t)\|_*}{8\beta_f}\}.$$

Proof. By the optimality of the point p_t we have that

$$\begin{aligned}
(p_t - x_t)\cdot\nabla f(x_t) &\leq (x^* - x_t)\cdot\nabla f(x_t) \\
&\leq f(x^*) - f(x_t) = -h_t, \qquad (4)
\end{aligned}$$

where the second inequality follows from convexity of f. Denote $c_t = \frac{1}{2}(x_t + p_t)$ and $w_t \in \arg\min_{w\in\mathbf{E},\|w\|\leq 1} w\cdot\nabla f(x_t)$. Note that from Holder's inequality we have that $w_t\cdot\nabla f(x_t) = -\|\nabla f(x_t)\|_*$. Using the strong convexity of the set $\mathcal{K}$ we have that the point $\tilde{p}_t = c_t + \frac{\alpha_K}{8}\|x_t - p_t\|^2 w_t$ is in $\mathcal{K}$. Again using the optimality of p_t we have that

$$\begin{aligned}
&(p_t - x_t)\cdot\nabla f(x_t) \leq (\tilde{p}_t - x_t)\cdot\nabla f(x_t) \\
&= \frac{1}{2}(p_t - x_t)\cdot\nabla f(x_t) + \frac{\alpha_K\|x_t - p_t\|^2}{8} w_t\cdot\nabla f(x_t) \\
&\leq -\frac{1}{2}h_t - \frac{\alpha_K\|x_t - p_t\|^2}{8}\|\nabla f(x_t)\|_*, \qquad (5)
\end{aligned}$$

where the last inequality follows from Eq. (4).

We now analyze the decrease in the approximation error h_{t+1}. By smoothness of f we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \eta_t(p_t - x_t)\cdot\nabla f(x_t) \\ &+ \frac{\beta_f}{2}\eta_t^2\|p_t - x_t\|^2. \end{aligned}$$

Subtracting $f(x^*)$ from both sides we have

$$h_{t+1} \leq h_t + \eta_t(p_t - x_t)\cdot\nabla f(x_t) + \frac{\beta_f}{2}\eta_t^2\|p_t - x_t\|^2. \tag{6}$$

Plugging Eq. (5) we have

$$\begin{aligned} h_{t+1} &\leq h_t\left(1 - \frac{\eta_t}{2}\right) - \eta_t\frac{\alpha_K\|x_t - p_t\|^2}{8}\|\nabla f(x_t)\|_* \\ &+ \frac{\beta_f}{2}\eta_t^2\|p_t - x_t\|^2 \\ &= h_t\left(1 - \frac{\eta_t}{2}\right) \\ &+ \frac{\|x_t - p_t\|^2}{2}\left(\eta_t^2\beta_f - \eta_t\frac{\alpha_K\|\nabla f(x_t)\|_*}{4}\right). \end{aligned}$$

In case $\frac{\alpha_K\|\nabla f(x_t)\|_*}{4} \geq \beta_f$, by the optimal choice of η_t in Algorithm 1, we can set $\eta_t = 1$ and get

$$h_{t+1} \leq \frac{h_t}{2}.$$

Otherwise, we can set $\eta_t = \frac{\alpha_K\|\nabla f(x_t)\|_*}{4\beta_f}$ and get

$$h_{t+1} \leq h_t\left(1 - \frac{\alpha_K\|\nabla f(x_t)\|_*}{8\beta_f}\right).$$

□

Note that Lemma 1 only relies on the strong convexity of the set $\mathcal{K}$ and did not assume anything regrading f beyond convexity and smoothness. In particular it does not require f to be strongly convex.

We can now prove Theorem 2.

Proof. Let $M = \frac{\sqrt{\alpha_f}\alpha_K}{8\sqrt{2}\beta_f}$ and $C = \max\{\frac{9}{2}\beta_f D_{\mathcal{K}}^2, 18M^{-2}\}$. We prove by induction that for all $t \geq 1$, $h_t \leq \frac{C}{(t+2)^2}$.

Since we assume that the objective function f satisfies Eq. (2), we have from Lemma 1 that on any iteration t,

$$\begin{aligned} h_{t+1} &\leq h_t\cdot\max\{\frac{1}{2}, 1 - \frac{\alpha_K\sqrt{\alpha_f}}{8\sqrt{2}\beta_f}\sqrt{h_t}\} \\ &= h_t\cdot\max\{\frac{1}{2}, 1 - Mh_t^{1/2}\}. \end{aligned} \tag{7}$$

For the base case $t = 1$ we need to prove that $h_1 = f(x_1) - f(x^*) \leq C/4$. By β_f smoothness of f we have

$$\begin{aligned} f(x_1) - f(x^*) &= f(x_0 + \eta_0(p_0 - x_0)) - f(x^*) \\ &\leq h_0 + \eta_0(p_0 - x_0)\cdot\nabla f(x_0) + \frac{\beta_f\eta_0^2}{2}D_{\mathcal{K}}^2 \\ &\leq h_0(1 - \eta_0) + \frac{\beta_f\eta_0^2}{2}D_{\mathcal{K}}^2, \end{aligned}$$

where the last inequality follows from convexity of f. By the optimal choice of η_0 we can in particular set $\eta_0 = 1$ which gives $h_1 \leq \frac{\beta_f}{2}D_{\mathcal{K}}^2 \leq C/9$.

Assume now that the induction holds for time $t \geq 1$, that is $h_t \leq \frac{C}{(t+2)^2}$.

If the result of the $\max$ operation in Eq. (7) is the first argument, that is $1/2$, we have that

$$\begin{aligned} h_{t+1} &\leq \frac{h_t}{2} \leq \frac{C}{2(t+2)^2} = \frac{C}{(t+3)^2}\cdot\frac{(t+3)^2}{2(t+2)^2} \\ &\leq \frac{C}{(t+3)^2}. \end{aligned} \tag{8}$$

where the last inequality holds for any $t \geq 1$.

We now turn to the case in which the result of the $\max$ operation in Eq. (7) is the second argument. We consider two cases.

If $h_t \leq \frac{C}{2(t+2)^2}$ then as in Eq. (8) it holds for any $t \geq 1$ that

$$h_{t+1} \leq h_t \leq \frac{C}{2(t+2)^2} \leq \frac{C}{(t+3)^2},$$

where the first inequality follows from Eq. (7).

Otherwise, $h_t > \frac{C}{2(t+2)^2}$. By Eq. (7) and the induction assumption we have

$$\begin{aligned} h_{t+1} &\leq h_t\left(1 - Mh_t^{1/2}\right) \\ &< \frac{C}{(t+2)^2}\left(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\right) \\ &= \frac{C}{(t+3)^2}\cdot\frac{(t+3)^2}{(t+2)^2}\left(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\right) \\ &= \frac{C}{(t+3)^2}\cdot\frac{(t+2)^2 + 2t + 5}{(t+2)^2}\left(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\right) \\ &< \frac{C}{(t+3)^2}\left(1 + \frac{3(t+2)}{(t+2)^2}\right)\left(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\right) \\ &= \frac{C}{(t+3)^2}\left(1 + \frac{3}{t+2}\right)\left(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\right). \end{aligned}$$

Thus for $C \geq \frac{18}{M^2}$ we have that

$$h_{t+1} \leq \frac{C}{(t+3)^2}\left(1+\frac{3}{t+2}\right)\left(1-\frac{3}{t+2}\right) < \frac{C}{(t+3)^2}.$$

□

4. Derivation of Previous Fast Rates Results and Extensions

4.1. Deriving the Linear Rate of Polayk & Levitin

Polyak & Levitin considered in (Levitin & Polyak, 1966) the case in which the feasible set is strongly convex, the objective function is smooth and there exists a constant $g > 0$ such that

$$\forall x \in \mathcal{K}: \quad \|\nabla f(x)\|_* \geq g. \tag{9}$$

They showed that under the lower-bounded gradient assumption, Algorithm 1 converges with a linear rate, that is $e^{-\Theta(t)}$. Clearly by plugging Eq. (9) into Lemma 1 we have that on each iteration t

$$h_{t+1} \leq h_t \cdot \max\{\frac{1}{2}, 1 - \frac{\alpha_k g}{8\beta_f}\}.$$

which results in the same exponentially fast convergence rate as in (Levitin & Polyak, 1966) and following works such as (Demyanov & Rubinov, 1970; Dunn, 1979).

4.2. Deriving a Linear Rate for Arbitrary Convex Sets in case x^* is in the Interior of the Set

Assume now that the feasible set $\mathcal{K}$ is convex but not necessarily strongly convex. We assume that the objective function f is smooth, convex, satisfies Eq. (2) with some constant α_f and admits a minimizer (not necessarily unique) x^* that lies in the interior of $\mathcal{K}$, i.e. there exists a parameter $r > 0$ such that the ball of radius r with respect to norm $\|\cdot\|$ centered at x^* is fully contained in $\mathcal{K}$ [2]. GuéLat and Marcotte (GuéLat & Marcotte, 1986) showed the under the above conditions, the Frank-Wolfe algorithm converges with a linear rate. We now show how a slight modification in the proof of Lemma 1 yields this linear convergence result.

Let w_t be as in the proof of Lemma 1, that is $w_t \in \arg\min_{w \in \mathbf{E}, \|w\| \leq 1} w \cdot \nabla f(x_t)$. Instead of defining the point $\tilde{p}_t$ as in the proof of Lemma 1 we define it to be $\tilde{p}_t = x^* + r w_t$. Because of our assumption on the location of x^*, it holds that $\tilde{p}_t \in \mathcal{K}$. We thus have that

$$(\tilde{p}_t - x_t) \cdot \nabla f(x_t) = (x_t^* - x_t) \cdot \nabla f(x_t) + r w_t \cdot \nabla f(x_t) \leq -r\|\nabla f(x_t)\|_*.$$

Plugging this into Eq. (6) we have

$$\begin{aligned} h_{t+1} &\leq h_t - \eta_t r \|\nabla f(x_t)\|_* + \frac{\beta_f \eta_t^2 D_{\mathcal{K}}^2}{2} \\ &\leq h_t - \eta_t r \sqrt{\frac{\alpha_f}{2}} \sqrt{h_t} + \frac{\beta_f \eta_t^2 D_{\mathcal{K}}^2}{2}. \end{aligned}$$

where $D_{\mathcal{K}}$ denotes the diameter of $\mathcal{K}$ with respect to norm $\|\cdot\|$ and the second inequality follows from Eq. (2). By the optimal choice of η_t, we can set $\eta_t = \frac{r\sqrt{\alpha_f}\sqrt{h_t}}{\sqrt{2}\beta_f D_{\mathcal{K}}^2}$ and get

$$h_{t+1} \leq h_t - \frac{r^2 \alpha_f}{4\beta_f D_{\mathcal{K}}^2} h_t,$$

which results in a linear convergence result.

4.3. Relaxing the Strong Convexity of f

So far we have considered the case in which the objective function f is strongly convex. Notice however that our main instrument for proving the accelerated convergence rate, i.e. Lemma 1, did not rely directly on strong convexity of f, but on the magnitude of the gradient, $\|\nabla f(x_t)\|_*$. We have seen in Eq. (2) that indeed if f is strongly convex than the gradient is at least of the order of $\sqrt{h_t}$. We now show that there exists functions which are not strongly convex but still satisfy Eq. (2) and hence our results apply also for them.

Consider the function

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2.$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Assume that $m < n$ and all rows of A are linearly independent. In this case the optimization problem $\min_{x \in \mathcal{K}} f(x)$ is the problem of finding a point in $\mathcal{K}$ that best satisfies an under-determined linear system in terms of the mean square error. An application of the Frank-Wolfe method to this problem was studied in (Beck & Teboulle, 2004). Under these assumptions, the function f is smooth and convex but not strongly convex since the Hessian matrix given by $A^\top A$ is not positive definite (note however that the matrix $AA^\top$ is positive definite).

The gradient of f is given by

$$\nabla f(x) = A^\top (Ax - b).$$

[2] We assume here that $\mathcal{K}$ is full-dimensional. In any other case, we can assume instead that the intersection of the ball centered at x^* with the smallest subspace containing $\mathcal{K}$ is fully contained in $\mathcal{K}$. In this case we also need to replace the gradient $\nabla f(x)$ with its projection onto this subspace, see also footnote 1.

Thus we have that

$$\begin{aligned}\|\nabla f(x)\|_2^2 &= [A^\top(Ax-b)]^\top[A^\top(Ax-b)] \\ &\geq \lambda_{\min}(AA^\top)\|Ax-b\|_2^2 \\ &\geq 2\lambda_{\min}(AA^\top)\Big(\frac{1}{2}\|Ax-b\|_2^2 \\ &\quad - \frac{1}{2}\|Ax^*-b\|_2^2\Big),\end{aligned}$$

where $\lambda_{\min}(AA^\top)$ denotes the smallest eigenvalue of $AA^\top$. Since $AA^\top$ is positive definite, $\lambda_{\min}(AA^\top) > 0$ and it follows that f satisfies Eq. (2).

Combining the result of this subsection with the previous one yields the linear convergence rate of the Frank-Wolfe method applied to the convex feasibility problem studied in (Beck & Teboulle, 2004).

5. Examples of Strongly Convex Sets

In this section we explore convex sets for which Theorem 2 is applicable. That is, convex sets which on one hand are strongly convex, and on the other, admit a simple and efficient implementation of a linear optimization oracle. We show that various norms that give rise to natural regularization functions in machine learning, induce convex sets that fit both of the above requirements. A summary of our findings is given in Table 5. We note that in all cases in which the norm parameter p is smaller than 2 (or one of the parameters s, p in case of group norms), we are not aware of a practical algorithm for computing the projection.

5.1. Partial Characterization of Strongly Convex Sets

The following lemma is taken from (Journée et al., 2010) (Theorem 12).

Lemma 2. *Let $\boldsymbol{E}$ be a finite vector space and let $f : \boldsymbol{E} \to \mathbb{R}$ be non-negative, α-strongly convex and β-smooth. Then the set $\mathcal{K} = \{x \mid f(x) \leq r\}$ is $\frac{\alpha}{\sqrt{2\beta r}}$-strongly convex.*

This lemma for instance shows that the Euclidean ball of radius r is $1/r$-strongly convex (by applying the lemma with $f = \|x\|_2^2$).

The following lemma will be useful to prove that convex sets that are induced by certain norms, which do not correspond to a smooth function as in the previous lemma, are strongly convex. The proof is given in the appendix.

Lemma 3. *Let $\boldsymbol{E}$ be a finite vector space, let $\|\cdot\|$ be a norm over $\boldsymbol{E}$ and assume that the function $f(x) = \|x\|^2$ is α-strongly convex over $\boldsymbol{E}$ with respect to the norm $\|\cdot\|$. Then for any $r > 0$, the set $\mathbb{B}_{\|\cdot\|}(r) = \{x \in \boldsymbol{E} \mid \|x\| \leq r\}$ is $\frac{\alpha}{2r}$-strongly convex with respect to $\|\cdot\|$.*

5.2. ℓ_p Balls for $p \in (1, 2]$

Given a parameter $p \geq 1$, consider the ℓ_p ball of radius r,

$$\mathbb{B}_p(r) = \{x \in \mathbb{R}^n \mid \|x\|_p \leq r\}.$$

The following lemma is proved in (Shwartz, 2007).

Lemma 4. *Fix $p \in (1, 2]$. The function $\frac{1}{2}\|x\|_p^2$ is $(p-1)$-strongly convex w.r.t. the norm $\|\cdot\|_p$.*

The following corollary is a consequence of combining Lemma 4 and Lemma 3. The proof is given in the appendix

Corollary 1. *Fix $p \in (1, 2]$. The set $\mathbb{B}_p(r)$ is $\frac{p-1}{r}$-strongly convex with respect to the norm $\|\cdot\|_p$ and $\frac{(p-1)n^{\frac{1}{2}-\frac{1}{p}}}{r}$-strongly convex with respect to the norm $\|\cdot\|_2$.*

The following lemma establishes that linear optimization over $\mathbb{B}_p(r)$ admits a simple closed-form solution that can be computed in time that is linear in the number of non-zeros in the linear objective. The proof is given in the appendix.

Lemma 5. *Fix $p \in (1, 2]$, $r > 0$ and a linear objective $c \in \mathbb{R}^n$. Let $x \in \mathbb{R}^n$ such that $x_i = -\frac{r}{\|c\|_q^{q-1}} sgn(c_i)|c_i|^{q-1}$ where q satisfies: $1/q + 1/p = 1$, and $sgn(\cdot)$ is the sign function. Then $x = \arg\min_{y \in \mathbb{B}_p(r)} y \cdot c$*

5.3. Schatten ℓ_p Balls for $p \in (1, 2]$

Given a matrix $X \in \mathbb{R}^{m \times n}$ we denote by $\sigma(X)$ the vector of singular values of X in descending order, that is $\sigma(X)_1 \geq \sigma(X)_2 \geq ...\sigma(X)_{\min(m,n)}$. The Schatten ℓ_p norm is given by

$$\|X\|_{S(p)} = \|\sigma(X)\|_p = \left(\sum_{i=1}^{\min(m,n)} \sigma(X)_i^p\right)^{1/p}.$$

Consider the Schatten ℓ_p ball of radius r,

$$\mathbb{B}_{S(p)}(r) = \{X \in \mathbb{R}^{m \times n} \mid \|X\|_{S(p)} \leq r\}.$$

The following lemma is taken from (Kakade et al., 2012).

Lemma 6. *Fix $p \in (1, 2]$. The function $\frac{1}{2}\|X\|_{S(p)}^2$ is $(p-1)$-strongly convex w.r.t. the norm $\|\cdot\|_{S(p)}$.*

The proof of the following corollary follows the exact same lines as the proof of Corollary 1 by using Lemma 6 instead of Lemma 4.

Corollary 2. *Fix $p \in (1, 2]$. The set $\mathbb{B}_{S(p)}(r)$ is $\frac{p-1}{r}$-strongly convex with respect to the norm $\|\cdot\|_{S(p)}$ and $\frac{(p-1)\min(m,n)^{\frac{1}{2}-\frac{1}{p}}}{r}$-strongly convex with respect to the frobenius norm $\|\cdot\|_F$.*

E	Domain name	Domain expression	S.C. parameter	Complexity of lin. opt.
$\mathbb{R}^n$	ℓ_p ball, $p \in (1,2]$	$\{x \in \mathbb{R}^n \mid \|x\|_p \leq r\}$	$\frac{p-1}{r}$	$O(\text{nnz})$
$\mathbb{R}^{m\times n}$	Schatten ℓ_p ball, $p \in (1,2]$	$\{X \in \mathbb{R}^{m\times n} \mid \|\sigma(X)\|_p \leq r\}$	$\frac{p-1}{r}$	$O(n^3)$ (SVD)
$\mathbb{R}^{m\times n}$	Group $\ell_{s,p}$ ball, $s,p \in (1,2]$	$\{X \in \mathbb{R}^{m\times n} \mid \|X\|_{s,p} \leq r\}$	$\frac{(s-1)(p-1)}{(s+p-2)r}$	$O(\text{nnz})$

Table 2. Examples of strongly convex sets with corresponding strong convexity parameter and complexity of a linear optimization oracle implementation . nnz denotes the number of non-zero entries in the linear objective and $\sigma(X)$ denotes the vector of singular values.

The following lemma establishes that linear optimization over $\mathbb{B}_{S(p)}(r)$ admits a simple closed-form solution given the *singular value decomposition* of the linear objective. The proof is given in the appendix.

Lemma 7. *Fix $p \in (1,2]$, $r > 0$ and a linear objective $C \in \mathbb{R}^{m\times n}$. Let $C = U\Sigma V^\top$ be the singular value decomposition of C. Let σ be a vector such that $\sigma_i = -\frac{r}{\|\sigma(C)\|_q^{q-1}}\sigma(C)_i^{q-1}$ where q satisfies: $1/q + 1/p = 1$. Finally, let $X = UDiag(\sigma)V^\top$ where $Diag(\sigma)$ is an $m \times n$ diagonal matrix with the vector σ as the main diagonal. Then $X = \arg\min_{Y\in\mathbb{B}_{S(p)}(r)} Y \bullet C$, where $\bullet$ denotes the standard inner product for matrices.*

5.4. Group $\ell_{s,p}$ Balls for $s,p \in (1,2]$

Given a matrix $X \in \mathbb{R}^{m\times n}$ denote by $X_i \in \mathbb{R}^n$ the ith row of X. That is $X = (X_1, X_2, ..., X_m)^\top$.

The $\ell_{s,p}$ norm of X is given by,

$$\|X\|_{s,p} = \|(\|X_1\|_s, \|X_2\|_s, ..., \|X_m\|_s)\|_p.$$

We define the $\ell_{s,p}$ ball as follows:

$$\mathbb{B}_{s,p}(r) = \{X \in \mathbb{R}^{m\times n} \mid \|X\|_{s,p} \leq r\}.$$

The proof of the following lemma is given in the appendix.

Lemma 8. *Let $s,p \in (1,2]$. The set $\mathbb{B}_{s,p}(r)$ is $\frac{(s-1)(p-1)}{(s+p-2)r}$-strongly convex with respect to the norm $\|\cdot\|_{s,p}$ and $n^{\frac{1}{s}-\frac{1}{2}}m^{\frac{1}{p}-\frac{1}{2}}\frac{(s-1)(p-1)}{(s+p-2)r}$-strongly convex with respect to the frobenius norm $\|\cdot\|_F$.*

The following lemma establishes that linear optimization over $\mathbb{B}_{s,p}(r)$ admits a simple closed-form solution that can be computed in time that is linear in the number of non-zeros in the linear objective. The proof is given in the appendix.

Lemma 9. *Fix $s,p \in (1,2]$, $r > 0$ and a linear objective $C \in \mathbb{R}^{m\times n}$. Let $X \in \mathbb{R}^{m\times n}$ be such that $X_{i,j} = -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}} sgn(C_{i,j})|C_{i,j}|^{z-1}$ where z satisfies: $1/s + 1/z = 1$, q satisfies: $1/p + 1/q = 1$ and C_i denotes the ith row of C. Then $X = \arg\min_{Y\in\mathbb{B}_{s,p}(r)} Y \bullet C$, where $\bullet$ denotes the standard inner product for matrices.*

6. Conclusions and Open Problems

In this paper we proved that the Frank-Wolfe algorithm converges at an accelerated rate of $O(1/t^2)$ for smooth and strongly-convex optimization over strongly-convex sets, beating the known tight convergence rate of the method for general smooth and convex optimization. This is one of the very few known results that achieve such an improvement in convergence rate under natural and standard assumptions (i.e. strong convexity etc.). We have further demonstrated that various regularization functions in machine learning give rise to strongly convex sets. We have also demonstrated how previous fast convergence rate results follow easily from our analysis.

The following questions naturally arise.

It is known that in case the objective function is both smooth and strongly convex, projection/prox-based methods achieve a convergence rate of $O(\log(1/\epsilon))$. Is it possible to achieve such a rate for the FW method in case the convex set is strongly convex?

We have shown that it is possible to obtain faster rates for optimization over balls induced by norms that give rise to strongly convex functions. Is it possible to obtain faster rates for balls induced by norms that do not give rise to strongly convex functions (but rather to smooth functions)? e.g. is it possible to obtain faster rates for ℓ_p balls for $p > 2$.

Finally, the most intriguing question is to give a linear optimization oracle-based method that enjoys the same convergence rate, at least in terms of the approximation error, as optimal projection/prox-based gradient methods, in any regime (including non-smooth problems). A progress in this direction was made recently by Garber and Hazan (Garber & Hazan, 2013b) who showed that in case the feasible set is a polytope, a variant of the FW-method obtains the same rates as the projected (sub)gradient descent method.

Acknowledgments

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 336078 – ERC-SUBLRN.

References

Ahipasaoglu, S. Damla, Sun, Peng, and Todd, Michael J. Linear convergence of a modified frank-wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.

Beck, Amir and Teboulle, Marc. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Meth. of OR*, 59(2):235–247, 2004.

Clarkson, Kenneth L. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2008.

Demyanov, Vladimir F. and Rubinov, Aleksandr M. *Approximate methods in optimization problems*. Elsevier Publishing Company, 1970.

Dudík, Miroslav, Harchaoui, Zaïd, and Malick, Jérôme. Lifted coordinate descent for learning with trace-norm regularization. *Journal of Machine Learning Research - Proceedings Track*, 22:327–336, 2012.

Dunn, Joseph C. Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals. *SIAM Journal on Control and Optimization*, 17 (2), 1979.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:149–154, 1956.

Garber, Dan and Hazan, Elad. Playing non-linear games with linear oracles. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, 2013a.

Garber, Dan and Hazan, Elad. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *CoRR*, abs/1301.4666, 2013b.

GuéLat, Jacques and Marcotte, Patrice. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35 (1), 1986.

Harchaoui, Zaïd, Douze, Matthijs, Paulin, Mattis, Dudík, Miroslav, and Malick, Jérôme. Large-scale image classification with trace-norm regularization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012.

Hazan, Elad. Sparse approximate solutions to semidefinite programs. In *8th Latin American Theoretical Informatics Symposium, LATIN*, 2008.

Hazan, Elad and Kale, Satyen. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, 2013.

Jaggi, Martin and Sulovský, Marek. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning, ICML*, 2010.

Journée, Michel, Nesterov, Yurii, Richtárik, Peter, and Sepulchre, Rodolphe. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.

Kakade, Sham M., Shalev-Shwartz, Shai, and Tewari, Ambuj. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

Lacoste-Julien, Simon and Jaggi, Martin. An affine invariant linear convergence analysis for frank-wolfe algorithms. *CoRR*, abs/1312.7864, 2013.

Lacoste-Julien, Simon, Jaggi, Martin, Schmidt, Mark W., and Pletscher, Patrick. Block-coordinate frank-wolfe optimization for structural svms. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, 2013.

Lan, Guanghui. The complexity of large-scale convex programming under a linear optimization oracle. *CoRR*, abs/1309.5550, 2013.

Laue, Sören. A hybrid algorithm for convex semidefinite optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012.

Levitin, Evgeny S and Polyak, Boris T. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6:1–50, 1966.

Migdalas, Athanasios. A regularization of the frankwolfe method and unification of certain nonlinear programming methods. *Mathematical Programming*, 65:331–345, 1994.

Shalev-Shwartz, Shai, Gonen, Alon, and Shamir, Ohad. Large-scale convex minimization with a low-rank constraint. In *Proceedings of the 28th International Conference on Machine Learning, ICML*, 2011.

Shwartz, Shay Sahlev. Phd thesis. 2007.

A. Proof of Theorem 1

Proof. Fix an iteration t. By the β_f-smoothness of f we have that

$$\begin{aligned} h_{t+1} &= f(x_t + \eta_t(p_t - x_t)) - f(x^*) \\ &\leq f(x_t) - f(x^*) + \eta_t(p_t - x_t) \cdot \nabla f(x_t) \\ &+ \frac{\eta_t^2 \beta_f}{2} \|p_t - x_t\|^2 \\ &\leq h_t - \eta_t h_t + \frac{\eta_t^2 \beta_f D_{\mathcal{K}}^2}{2}, \end{aligned} \tag{10}$$

where the last inequality follows from convexity of f. Notice that by the optimal choice of η_t in Algorithm 1, it holds in particular that $h_{t+1} \leq h_t$ (by setting $\eta_t = 0$ in Eq. (10)).

Fix $C = 8\beta_f D_{\mathcal{K}}^2$. We now prove by induction on t that $h_t \leq \frac{C}{t}$.

For the base case $t = 1$ we notice that by the optimal choice of η_0 in Algorithm 1 we can in particular set $\eta_0 = 1$ and thus it follows from Eq. (10) that $h_1 \leq \frac{\beta_f D_{\mathcal{K}}^2}{2} < C$ as needed.

Assume now that the induction holds for $t \geq 1$. That is $h_t \leq \frac{C}{t}$. We consider two cases.

If $h_t \leq \frac{C}{2t}$ then we have

$$h_{t+1} \leq h_t \leq \frac{C}{2t} = \frac{C}{t+1} \cdot \frac{t+1}{2t} \leq \frac{C}{t+1},$$

where the last inequality holds for any $t \geq 1$.

Otherwise it holds that $h_t > \frac{C}{2t}$. Using Eq. (10) again we have

$$h_{t+1} \leq h_t - \eta_t h_t + \frac{\eta_t^2 \beta_f D_{\mathcal{K}}^2}{2}.$$

By the optimal choice of η_t in Algorithm 1 we can set $\eta_t = \frac{h_t}{\beta_f D_{\mathcal{K}}^2}$ and get

$$\begin{aligned} h_{t+1} &\leq h_t - \frac{1}{2\beta_f D_{\mathcal{K}}^2} h_t^2 < \frac{C}{t} - \frac{C^2}{8\beta_f D_{\mathcal{K}}^2 t^2} \\ &= \frac{C}{t+1}\left(\frac{t+1}{t} - \frac{C(t+1)}{8\beta_f D_{\mathcal{K}}^2 t^2}\right) \\ &< \frac{C}{t+1}\left(1 + \frac{1}{t} - \frac{Ct}{8\beta_f D_{\mathcal{K}}^2 t^2}\right). \end{aligned}$$

Thus for $C \geq 8\beta_f D_{\mathcal{K}}^2$ we have that $h_{t+1} \leq \frac{C}{t+1}$. □

B. Proofs of Lemmas and Corollaries from Section 5

B.1. Proof of Lemma 3

Proof. It suffices to show that given $x, y \in \mathbf{E}$ such that $f(x) \leq r^2, f(y) \leq r^2$, a scalar $\gamma \in [0,1]$ and $z \in \mathbf{E}$ such that $\|z\| \leq \frac{\alpha}{4r}\gamma(1-\gamma)\|x-y\|^2$, it holds that, $f(\gamma x + (1-\gamma)y + z) \leq r^2$.

By the definition of f and the triangle inequality for $\|\cdot\|$ we have

$$\begin{aligned} &f(\gamma x + (1-\gamma)y + z) = \|\gamma x + (1-\gamma)y + z\|^2 \leq \\ &\left(\|\gamma x + (1-\gamma)y\| + \|z\|\right)^2 = \\ &\left(\sqrt{f(\gamma x + (1-\gamma)y)} + \|z\|\right)^2. \end{aligned} \tag{11}$$

Since f is α strongly convex with respect to $\|\cdot\|$ we have that

$$\begin{aligned} &f(\gamma x + (1-\gamma)y) \leq \\ &\gamma f(x) + (1-\gamma)f(y) - \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|^2 \leq \\ &r^2 - \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|^2. \end{aligned}$$

The function $g(t) = \sqrt{t}$ is concave, meaning $\sqrt{a-b} = g(a-b) \leq g(a) - g'(a)\cdot b = \sqrt{a} - \frac{b}{2\sqrt{a}}$. Thus,

$$\begin{aligned} \sqrt{f(\gamma x + (1-\gamma)y)} &\leq \sqrt{r^2 - \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|^2} \\ &\leq r - \frac{\alpha\gamma(1-\gamma)\|x-y\|^2}{4r}. \end{aligned}$$

Plugging back in Eq. (11) we have

$$\begin{aligned} &f(\gamma x + (1-\gamma)y + z) \leq \\ &\left(r - \frac{\alpha\gamma(1-\gamma)\|x-y\|^2}{4r} + \|z\|\right)^2. \end{aligned}$$

By our assumption on $\|z\|$ we have

$$\begin{aligned} f(\gamma x + (1-\gamma)y + z) &\leq \left(r - \frac{\alpha\gamma(1-\gamma)\|x-y\|^2}{4r}\right. \\ &+ \left.\frac{\alpha}{4r}\gamma(1-\gamma)\|x-y\|^2\right)^2 \\ &= r^2. \end{aligned}$$

□

B.2. Proof of Corollary 1

Proof. The strong convexity of the set w.r.t. $\|\cdot\|_p$ is an immediate consequence of Lemma 3.

Since $\mathbb{B}_p(r)$ is $\alpha = (p-1)/r$ strongly convex w.r.t. the norm $\|\cdot\|_p$, we have that given $x, y \in \mathbb{B}_p(r)$, $\gamma \in [0,1]$ and $z \in \mathbb{R}^n$ such that $\|z\|_p \leq 1$ it holds that

$$\gamma x + (1-\gamma)y + \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_p^2 z \in \mathbb{B}_p(r).$$

For any $p \in (1,2]$ and vector $v \in \mathbb{R}^n$ it holds that

$$\|v\|_2 \leq \|v\|_p \leq n^{\frac{1}{p}-\frac{1}{2}}\|v\|_2. \tag{12}$$

Given a vector $z' \in \mathbb{R}^n$ such that $\|z'\|_F \leq 1$ we have that

$$\|\frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_2^2 z'\|_p = \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_2^2\|z'\|_p.$$

Using Eq. (12) we have

$$\|\frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_2^2 z'\|_p \leq \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_p^2 n^{\frac{1}{p}-\frac{1}{2}}\|z'\|_2 \leq \frac{\alpha n^{\frac{1}{p}-\frac{1}{2}}}{2}\gamma(1-\gamma)\|x-y\|_p^2.$$

Hence, $\mathbb{B}_p(r)$ is $\alpha n^{\frac{1}{2}-\frac{1}{p}} = \frac{(p-1)n^{\frac{1}{2}-\frac{1}{p}}}{r}$-strongly convex with respect to $\|\cdot\|_2$. □

B.3. Proof of Lemma 5

Proof. Since $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms, we have using Holder's inequality that for all $x \in \mathbb{B}_p(r)$,

$$x \cdot c \geq -\|x\|_p\|c\|_q \geq -r\|c\|_q.$$

Thus choosing $x_i = -\frac{r}{\|c\|_q^{q-1}}\text{sgn}(c_i)|c_i|^{q-1}$ we have that

$$\begin{aligned} x \cdot c &= -\sum_{i=1}^n \frac{r}{\|c\|_q^{q-1}}\text{sgn}(c_i)|c_i|^{q-1}\cdot c_i \\ &= -\sum_{i=1}^n \frac{r}{\|c\|_q^{q-1}}|c_i|^q = -\frac{r}{\|c\|_q^{q-1}}\|c\|_q^q \\ &= -r\|c\|_q. \end{aligned}$$

Moreover,

$$\|x\|_p^p = \frac{r^p}{\left(\|c\|_q^{q-1}\right)^p}\sum_{i=1}^n \left(|c_i|^{q-1}\right)^p.$$

Since $p = q/(q-1)$ we have that

$$\|x\|_p^p = \frac{r^p}{\|c\|_q^q}\sum_{i=1}^n |c_i|^q = r^p.$$

Thus we have that $x \in \mathbb{B}_p(r)$. □

B.4. Proof of Lemma 7

Proof. Since $\|\cdot\|_{S(p)}$ and $\|\cdot\|_{S(q)}$ are dual norms we from Holder's inequality that for all $X \in \mathbb{B}_{S(p)}(r)$,

$$X \bullet C \geq -\|X\|_{S(p)}\|C\|_{S(q)} \geq -r\|C\|_{S(q)} = r\|\sigma(C)\|_q.$$

By our choice of X we have that

$$\begin{aligned} X \bullet C &= \text{Tr}(X^\top C) = \text{Tr}(V Diag(\sigma)^\top U^\top U \Sigma V^\top) \\ &= \text{Tr}(V Diag(\sigma)^\top \Sigma V^\top) \\ &= \text{Tr}(V^\top V Diag(\sigma)^\top \Sigma) = \text{Tr}(Diag(\sigma)^\top \Sigma) \\ &= \sum_{i=1}^{\min(m,n)} -\frac{r}{\|\sigma(C)\|_q^{q-1}}\sigma(C)_i^{q-1}\cdot\sigma(C)_i \\ &= -\frac{r}{\|\sigma(C)\|_q^{q-1}}\sum_{i=1}^{\min(m,n)}\sigma(C)_i^q \\ &= -r\|\sigma(C)\|_q. \end{aligned}$$

Moreover,

$$\|X\|_{S(p)}^p = \|\sigma(X)\|_p^p = \frac{r^p}{\left(\|\sigma(C)\|_q^{q-1}\right)^p}\sum_{i=1}^n\left(\sigma(C)_i^{q-1}\right)^p.$$

Since $p = q/(q-1)$ we have that

$$\|X\|_{S(p)}^p = \frac{r^p}{\|\sigma(C)\|_q^q}\sum_{i=1}^n|\sigma(C)_i|^q = r^p.$$

Thus we have that $X \in \mathbb{B}_{S(p)}(r)$. □

B.5. Proof of Lemma 8

The following lemma will be of use in the proof.

Lemma 10. *for any matrix $A \in \mathbb{R}^{m\times n}$ and $s,p \in (1,2]$ it holds that*

$$\|A\|_F \leq \|A\|_{s,p} \leq n^{\frac{1}{s}-\frac{1}{2}}m^{\frac{1}{p}-\frac{1}{2}}\|A\|_F.$$

Proof. For any vector $v \in \mathbb{R}^n$ and $p \in (1,2]$ it holds that

$$\|v\|_2 \leq \|v\|_p \leq n^{\frac{1}{p}-\frac{1}{2}}\|v\|_2. \tag{13}$$

Denote by A_i the ith row of A. For any $i \in [m]$ and $p \in (1,2]$ it holds that

$$\|A_i\|_2 \leq \|A_i\|_p \leq n^{\frac{1}{p}-\frac{1}{2}}\|A_i\|_2. \tag{14}$$

Note that by definition $\|\cdot\|_F \equiv \|\cdot\|_{2,2}$. Applying Eq. (13) and (14) we have,

$$\begin{aligned} \|A\|_F &= \|A\|_{2,2} = \|(\|A_1\|_2, \|A_2\|_2, ..., \|A_m\|_2)\|_2 \\ &\leq \|(\|A_1\|_s, \|A_2\|_s, ..., \|A_m\|_s)\|_p \\ &\leq n^{\frac{1}{s}-\frac{1}{2}}m^{\frac{1}{p}-\frac{1}{2}}\|(\|A_1\|_2, \|A_2\|_2, ..., \|A_m\|_2)\|_2 \\ &= n^{\frac{1}{s}-\frac{1}{2}}m^{\frac{1}{p}-\frac{1}{2}}\|A\|_F. \end{aligned}$$

□

We can now prove Lemma 8.

Proof. Let z, q be such that $1/z+1/s = 1$ and $1/q+1/p = 1$. Note that $z, q \in [2, \infty)$. The norm $\|\cdot\|_{z,q}$ is the dual norm to $\|\cdot\|_{s,p}$ (see (Kakade et al., 2012) for instance).

According to Lemma 4, the functions $\|x\|_s^2$ and $\|x\|_p^2$ are $\alpha_s = 2(s-1)$-strongly convex w.r.t. $\|\cdot\|_p$ and $\alpha_p = 2(p-1)$-strongly convex w.r.t. $\|\cdot\|_q$ respectively. Hence by the *strong convexity / smoothness duality* (see Theorem 3 in (Kakade et al., 2012)) we have that the functions $\|x\|_z^2$ and $\|x\|_q^2$ are α_s^{-1}-smooth w.r.t. $\|\cdot\|_z$ and α_p^{-1}-smooth w.r.t. $\|\cdot\|_q$ respectively.

By Theorem 13 in (Kakade et al., 2012) we have that the function $\|X\|_{z,q}^2$ is $(\alpha_p^{-1}+\alpha_s^{-1})$-smooth with respect to the norm $\|\cdot\|_{z,q}$. Again using the *strong convexity / smoothness duality* we have that $\|X\|_{s,p}^2$ is $\left(\alpha_p^{-1}+\alpha_s^{-1}\right)^{-1} = \frac{\alpha_p\alpha_s}{\alpha_p+\alpha_s}$ strongly convex with respect to the norm $\|\cdot\|_{s,p}$. The first part of the lemma now follows from applying Lemma 3.

Since $\mathbb{B}_{s,p}(r)$ is $\alpha = \frac{(s-1)(p-1)}{(s+p-2)r}$ strongly convex w.r.t. the norm $\|\cdot\|_{s,p}$, we have that given $X, Y \in \mathbb{B}_{s,p}(r)$, $\gamma \in [0,1]$ and $Z \in \mathbb{R}^{m\times n}$ such that $\|Z\|_{s,p} \le 1$ it holds that

$$\gamma X + (1-\gamma)Y + \frac{\alpha}{2}\gamma(1-\gamma)\|X-Y\|_{s,p}^2 Z \in \mathbb{B}_{s,p}(r).$$

Given a matrix $Z' \in \mathbb{R}^{m\times n}$ such that $\|Z'\|_F \le 1$ we have that

$$\|\frac{\alpha}{2}\gamma(1-\gamma)\|X-Y\|_F^2 Z'\|_{s,p} = \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_F^2\|Z'\|_{s,p}.$$

Using Lemma 10 we have

$$\|\frac{\alpha}{2}\gamma(1-\gamma)\|X-Y\|_F^2 Z'\|_{s,p} \le \frac{\alpha}{2}\gamma(1-\gamma)\|X-Y\|_{s,p}^2 n^{\frac{1}{s}-\frac{1}{2}} m^{\frac{1}{p}-\frac{1}{2}}\|Z'\|_F \le \frac{\alpha n^{\frac{1}{s}-\frac{1}{2}} m^{\frac{1}{p}-\frac{1}{2}}}{2}\gamma(1-\gamma)\|X-Y\|_{s,p}^2.$$

Hence, $\mathbb{B}_{s,p}(r)$ is $\alpha n^{\frac{1}{s}-\frac{1}{2}} m^{\frac{1}{p}-\frac{1}{2}} = n^{\frac{1}{s}-\frac{1}{2}} m^{\frac{1}{p}-\frac{1}{2}} \frac{(s-1)(p-1)}{(s+p-2)r}$ strongly convex with respect to $\|\cdot\|_F$. □

B.6. Proof of Lemma 9

Proof. Since by choice of z, q it holds that $\|\cdot\|_{s,p}, \|\cdot\|_{z,q}$ are dual norms, we have by Holder's inequality that

$$X \bullet C \ge -\|X\|_{s,p}\|C\|_{z,q} \ge -r\|C\|_{z,q}.$$

Thus, choosing

$$X_{i,j} = -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}}\text{sgn}(C_{i,j})|C_{i,j}|^{z-1},$$

we have that

$$X \bullet C = \sum_{i\in[m],j\in[n]} X_{i,j}C_{i,j} = \sum_{i\in[m],j\in[n]} -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}}\text{sgn}(C_{i,j})|C_{i,j}|^{z-1}\cdot C_{i,j} = \sum_{i\in[m],j\in[n]} -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}}|C_{i,j}|^z = \sum_{i\in[m]} -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}}\sum_{j\in[n]}|C_{i,j}|^z = \sum_{i\in[m]} -\frac{r}{\|C\|_{z,q}^{q-1}\|C_i\|_z^{z-q}}\|C_i\|_z^z = \sum_{i\in[m]} -\frac{r}{\|C\|_{z,q}^{q-1}}\|C_i\|_z^q = -\frac{r}{\|C\|_{z,q}^{q-1}}\sum_{i\in[m]}\|C_i\|_z^q = -\frac{r}{\|C\|_{z,q}^{q-1}}\|C\|_{z,q}^q = -r\|C\|_{z,q}.$$

Moreover, for all $i \in [m]$ it holds that

$$\|X_i\|_s^s = \sum_{j=1}^n |X_{i,j}|^s = \frac{r^s}{\|C\|_{z,q}^{s(q-1)}\|C_i\|_z^{s(z-q)}}\sum_{i=j}^n |C_{i,j}|^{s(z-1)}.$$

Since $s = z/(z-1)$ we have

$$\|X_i\|_s^s = \frac{r^s}{\|C\|_{z,q}^{s(q-1)}\|C_i\|_z^{s(z-q)}}\|C_i\|_z^z = \frac{\|C_i\|_z^{sq-z(s-1)}}{\|C\|_{z,q}^{s(q-1)}}r^s$$

Using $z = s/(s-1)$ we have that

$$\|X_i\|_s^s = \frac{\|C_i\|_z^{s(q-1)}}{\|C\|_{z,q}^{s(q-1)}}r^s.$$

Thus,

$$\|X_i\|_s = \left(\frac{\|C_i\|_z}{\|C\|_{z,q}}\right)^{q-1} r.$$

Finally, we have that

$$\|X\|_{s,p}^p = \sum_{i\in[m]}\|X_i\|_s^p = \sum_{i\in[m]}\left(\frac{\|C_i\|_z}{\|C\|_{z,q}}\right)^{p(q-1)} r^p = \sum_{i\in[m]}\left(\frac{\|C_i\|_z}{\|C\|_{z,q}}\right)^q r^p = \frac{r^p}{\|C\|_{z,q}^q}\sum_{i\in[m]}\|C_i\|_z^q = r^p.$$

Thus, $X \in \mathbb{B}_{s,p}(r)$. □