Polytope conditioning and linear convergence of the Frank-Wolfe algorithm

Javier Peña* Daniel Rodríguez†

December 24, 2016

Abstract

It is known that the gradient descent algorithm converges linearly when applied to a strongly convex function with Lipschitz gradient. In this case the algorithm's rate of convergence is determined by the condition number of the function. In a similar vein, it has been shown that a variant of the Frank-Wolfe algorithm with away steps converges linearly when applied to a strongly convex function with Lipschitz gradient over a polytope. In a nice extension of the unconstrained case, the algorithm's rate of convergence is determined by the product of the condition number of the function and a certain condition number of the polytope.

We shed new light into the latter type of polytope conditioning. In particular, we show that previous and seemingly different approaches to define a suitable condition measure for the polytope are essentially equivalent to each other. Perhaps more interesting, they can all be unified via a parameter of the polytope that formalizes a key premise linked to the algorithm's linear convergence. We also give new insight into the linear convergence property. For a convex quadratic objective, we show that the rate of convergence is determined by a condition number of a suitably scaled polytope.

*Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu

†Department of Mathematical Sciences, Carnegie Mellon University, USA, drod@cmu.edu

1 Introduction

It is a standard result in convex optimization that the gradient descent algorithm converges linearly to the minimizer of a strongly convex function with Lipschitz gradient. For a related discussion, e.g., [13, Chapter 2] or [4, Chapter 1]. Furthermore, in this case the rate of convergence is determined by the *condition number* of the objective function, that is, the ratio between the Lipschitz parameter of the gradient and the strong convexity parameter of the function.

In analogous fashion, the Frank-Wolfe algorithm [6, 9], also known as conditional gradient algorithm, for the problem $\min_{u \in C} f(u)$ converges linearly to the minimizer of f on a compact convex set C provided f is strongly convex with Lipschitz gradient and the optimal solution lies in the relative interior of C. For a related discussion see, e.g., [3, 5, 8, 11] and the references therein. The assumption that the optimal solution belongs to relative interior of C is critical for the linear convergence of the algorithm. Indeed, the rate of convergence depends on how far the optimal solution is from the relative boundary of C. In particular, this rate deteriorates when the optimal solution is near the relative boundary of C, and linear convergence breaks down altogether when the optimal solution is on the relative boundary of C.

A variant of the Frank-Wolfe algorithm that includes *away steps* was proposed by Wolfe [15]. Several articles have shown linear convergence results for the away step variant and for other variants of the Frank-Wolfe algorithm when the domain C is a polyhedron. The article [8] establishes a local linear convergence result for the away step variant for strongly convex with Lipschitz gradient under a certain kind of strict complementarity assumption. The articles [1, 10] give local linear convergence results for smooth convex functions by relying on Robinson's second-order constraint qualification. On the other hand, [12] shows linear convergence results for a pairwise variant of the Frank-Wolfe algorithm. The article [7] shows linear convergence for a version of the Frank-Wolfe algorithm that relies on a local linear optimization oracle.

The recent articles [2, 11, 14] establish *global* linear convergence results for the Frank-Wolfe algorithm with away steps when the objective function is strongly convex with Lipschitz gradient and the domain is of the form $C = \mathrm{conv}(A)$ for a finite set of atoms A. It should be noted that both [2] and [14] were inspired by and relied upon key ideas and results first introduced in a preliminary workshop version of [11]. A common feature

of [2, 11, 14] is that the statement of linear convergence is given in terms of the condition number of the objective function f and some type of *condition number* of the polytope $\mathrm{conv}(A)$. As we detail in Section 3, a generic version of linear convergence as in [2, 11, 14] hinges on three main premises. The first premise is a certain *slope bound* on the objective function and its optimal solution set. This first premise readily holds for strongly convex functions as it does in the unconstrained case. The second premise is a *decrease condition* on the objective function at each iteration of the algorithm. As in the unconstrained case, the second premise holds as long as an upper bound on the Lipschitz constant of the gradient is available, or if an appropriate line-search is performed at each iteration. The third premise, which seems to be the main technical component in each of the papers [2, 11, 14], is a premise on the search direction selected by the algorithm at each iteration. Loosely speaking, this third premise is a condition on the alignment of the search direction with the gradient of the objective function at the current iterate. The premise is that this alignment should be comparable to that of a direct step towards the optimal solution. Unlike the first two premises, that essentially match the premises leading to the linear convergence of gradient descent in the unconstrained case, the third premise is inherent to the polytope defining the constraint set of the problem. This third premise can be formalized in terms of a certain kind of *condition number* of the polytope. In a nice extension of the unconstrained case, the rate of linear convergence of the Frank-Wolfe algorithm with away steps is determined by the product of the usual condition number of the objective function and the condition number of the polytope. (See Theorem 4 in Section 3.)

The central goal of this paper is to shed new light into this polytope condition number. The three articles [2, 11, 14] make different attempts to define a suitable condition measure along the lines of the third premise sketched above. Each of these attempts has different merits and limitations. One of this paper's main contributions is to show that these three kinds of condition measures, namely the *pyramidal width* defined by Lacoste-Julien and Jaggi [11], the *vertex-facet distance* defined by Beck and Shtern [2], and the *restricted width* defined by Peña, Rodríguez, and Soheili [14], turn out to be essentially equivalent. Perhaps more important, they are all unified via a *facial distance* of the polytope. As we explain in Section 2 and Section 3, the facial distance can be seen as a natural quantity associated to the polytope that formalizes a key alignment condition of the search direction at each iteration of the Frank-Wolfe algorithm with away steps.

Section 2 presents the technical bulk of our paper. One of our results (Theorem 1) is a characterization of the facial distance of a polytope as

the minimum distance between a proper face of the polytope and a kind of *complement polytope*. This characterization can be seen as a refinement of the *vertex-facet distance* proposed by Beck and Shtern [2]. Theorem 1 motivates the name "facial distance" for this quantity. Theorem 1 provides a method to compute or bound the facial distance as we illustrate in a few examples. We also show (Theorem 2) that the facial distance coincides with the pyramidal width defined by Lacoste-Julien and Jaggi [11]. As a byproduct of this result, we obtain a simplification of the original definition of pyramidal width. We also give a *localized* version of Theorem 1 for a kind of *localized* version of the facial distance of the polytope (Theorem 3).

As mentioned above, Section 3 details how the linear convergence of the Frank-Wolfe algorithm with away steps can be derived from three central premises. The goal of Section 3 is to highlight the role of these three key premises, particularly the third one. We discuss how the third premise is naturally tied to the facial distance of the polytope discussed in Section 2. Our exposition allows us to distill a key tradeoff in the existing bounds on the rate of convergence of the Frank-Wolfe algorithm with away steps. On the one hand, the algorithm's rate of convergence can be bounded in terms of quantities that depend *only* on properties of the polytope and of the objective function but not on the optimal solution. More precisely, for a strongly convex objective function with Lipschitz gradient the rate of convergence can be bounded in terms of the product of the condition number of the polytope and the condition number of the objective function. This is a feature of the results in [2, 11] but not of those in [14] that depend on the optimal solution set. On the other hand, a *sharper* bound on the rate of convergence can be given if we allow it to depend on the location of the optimal solution in the polytope. More precisely, the rate of convergence can be bounded in terms of the product of a *localized* condition number of the polytope that depends on the solution set and the condition number of the objective function. The statement of Theorem 4 makes the connection between the two bounds completely transparent: The solution-independent bound is simply the most conservative solution-dependent one. Not surprisingly, the solution-dependent bound can be arbitrarily better than the solution-independent one.

Section 4 discusses the linear convergence property in the special but important case when the objective function is of the form $f(u) = \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle$ for Q positive semidefinite. As Theorem 5 in Section 4 shows, in this case the rate of convergence is determined by a variant of the facial distance of a suitably scaled polytope. In Section 5 we show that the latter result extends, under suitable assumptions on the algorithm's choice of steplength,

to a composite objective function of the form $f(u) = h(Eu) + \langle b, u\rangle$ where h is a strongly convex function with Lipschitz gradient and E is a matrix of suitable size. (See Theorem 6 in Section 5.) This result is along the lines of the linear convergence result of Beck and Shtern's [2, Theorem 3.1]. However, our bound on the rate of convergence and proof technique are fairly different. Both of them are extensions of the three-premise approach described in Section 3. We conclude our paper with some examples in Section 6 that illustrate the tightness of the linear convergence results stated in Theorem 4 and Theorem 5.

Throughout the paper we will often need to deal with multiple points in $\mathbb{R}^m$ and in $\mathbb{R}^n$. We will consistently use u, v to denote points in $\mathbb{R}^m$ and w, x, y, z to denote points in $\mathbb{R}^n$.

2 The facial distance $\Phi(A)$

This section constitutes the technical bulk of the paper. We define the *facial distance* $\Phi(A)$ and prove several interesting results about it. In particular, we show that it essentially matches the various kinds of condition measures previously defined in [2, 11, 14].

Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$. For convenience we will make the following slight abuse of notation: We will write A to denote both the matrix $\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$ and the set of its columns $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^m$. The appropriate meaning of A will be clear from the context. Let $\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$. For $x \in \Delta_{n-1}$, define $I(x) \subseteq \{1, \ldots, n\}$ and $S(x) \subseteq A$ as

$$I(x) := \{i \in \{1, \ldots, n\} : x_i > 0\}$$

and

$$S(x) := \{a_i : i \in I(x)\}.$$

Observe that the sets $I(x), S(x)$ define the *support* of Ax.

Throughout the paper, $\|\cdot\|$ will denote the Euclidean norm. Assume $x, z \in \Delta_{n-1}$ are such that $A(x - z) \neq 0$ and let $d := \frac{A(x-z)}{\|A(x-z)\|}$. Define

$$\Phi(A, x, z) = \min_{p\in\mathbb{R}^m : \langle p,d\rangle = 1} \max_{s\in S(x),\, a\in A} \langle p, s - a\rangle , \tag{1}$$

and

$$\Phi(A) = \min_{x,z\in\Delta_{n-1} : A(x-z)\neq 0} \Phi(A, x, z).$$

The connection between $\Phi(A, x, z)$ and the Frank-Wolfe algorithm with away steps algorithm will be made explicit as Premise 3 in Section 3 but

the basic idea is as follows. At each iteration the algorithm starts from a current point $u = Ax \in \text{conv}(A)$ and selects the two atoms $a, s \in A$ that attain $\max_{s \in S(x),\, a \in A} \langle p, s - a \rangle$ for $p = \nabla f(u)$. Premise 3 requires that for $d := \frac{A(x-z)}{\|A(x-z)\|}$ the ratio $\frac{\langle p, s-a \rangle}{\langle p, d \rangle}$ be bounded away from zero, where $z \in \Delta_{n-1}$ is such that $u^* = Az$ is a solution to the minimization problem. The latter condition means the alignment of the vector $a - s$ and the direction p should be comparable to the alignment of d and p. The need to formalize this premise motivates the definition of the quantities $\Phi(A, x, z)$ and $\Phi(A)$.

Notice the asymmetry between the roles x and z in $\Phi(A, x, z)$. We can think of z as defining an *anchor* point $Az \in \text{conv}(A)$. This anchor point determines a set of directions $d = \frac{A(x-z)}{\|A(x-z)\|}$ for $x \in \Delta_{n-1}$ with $A(x - z) \neq 0$. When $Az = 0$, the quantity $\Phi(A, x, z)$ coincides with the quantity $\phi(A, x)$ defined in [14]. Thus $\Phi(A)$ can be seen as a refinement of the *restricted width* $\phi(A) = \min_{x \in \Delta_{n-1} : Ax \neq 0} \phi(A, x)$ defined in [14]. When $0 \in \text{conv}(A)$, we have $\Phi(A) \leq \phi(A)$. More precisely, when $0 \in \text{conv}(A)$ the restricted width $\phi(A)$ coincides with the *localized* variant $\Phi(A, Z)$ of $\Phi(A)$ defined below for $Z = \{z \in \Delta_{n-1} : Az = 0\}$.

The quantity $\phi(A, x)$ was introduced in a form more closely related to the alternative expression (2) for $\Phi(A, x, z)$ in Proposition 1 below. The expression (2) characterizes $\Phi(A, x, z)$ as the length of the longest segment in $\text{conv}(A)$ in the direction $A(x - z)$ with one endpoint in $\text{conv}(S(x))$ and the other in $\text{conv}(A)$.

Proposition 1 *Assume* $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ *and* $x, z \in \Delta_{n-1}$ *are such that* $A(x - z) \neq 0$ *and let* $d := \frac{A(x-z)}{\|A(x-z)\|}$. *Then*

$$\Phi(A, x, z) = \max \{\lambda > 0 : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ A(w - y) = \lambda d\}\,. \tag{2}$$

Furthermore, if p *attains the minimum value* $\Phi(A, x, z)$ *in* (1) *and* $u = Aw, v = Ay$ *maximize the right hand side in* (2) *then* $v \in \text{conv}(B)$ *where* $B = \operatorname*{Argmin}_{a \in A} \langle p, a \rangle$, $u \in \text{conv}(A \setminus B)$, *and*

$$\Phi(A, x, z) = \|u - v\|.$$

Proof: To ease notation, let $I := I(x)$. Observe that the right-hand-side in

(2) is

$$\begin{array}{ll} \max\limits_{w_I,y,\lambda} & \lambda \\ & A_I w_I - Ay - \lambda d = 0 \\ & \mathbf{1}_I^{\mathrm{T}} w_I = 1 \\ & \mathbf{1}^{\mathrm{T}} y = 1 \\ & w_I, y \geq 0. \end{array}$$

On the other hand, from the definition (1) of $\Phi(A, x, z)$, it follows that

$$\begin{array}{lll} \Phi(A,x,z) = \min\limits_{p,t,\tau} & t + \tau \\ & A_I^{\mathrm{T}} p \leq t\mathbf{1}_I \\ & A^{\mathrm{T}} p \geq -\tau \mathbf{1} \\ & \langle d, p \rangle = 1. \end{array}$$

Therefore (2) follows by linear programming duality.

Next assume p attains the minimum value $\lambda = \Phi(A, x, z)$ in (1) and $u = Aw, v = Ay$ maximize the right hand side in (2). Then (w_I, y, λ) and (p, t, τ) are respectively solutions to the above pair of linear programs for $t = \max_{a \in A} \langle p, a \rangle$ and $-\tau = \min_{a \in A} \langle p, a \rangle$. By complementary slackness it follows that $y_i > 0$ only if $a_i^{\mathrm{T}} p = -\tau = \min_{a \in A} \langle p, a \rangle$. Thus $v = Ay \in \mathrm{conv}(B)$ for $B = \mathrm{Argmin}_{a \in A} \langle p, a \rangle$. Similarly, by complementary slackness it follows that $w_j > 0$ only if $a_j^{\mathrm{T}} p = t = \max_{a \in S(x)} \langle p, a \rangle$. Thus $u = Aw \in \mathrm{conv}(C)$ for $C = \mathrm{Argmax}_{a \in S(x)} \langle p, a \rangle$. Next observe that $C \subseteq A \setminus B$ because $\max_{a \in S(x)} \langle p, a \rangle - \min_{a \in A} \langle p, a \rangle = t + \tau = \lambda \geq \|A(x - z)\| > 0$. Finally observe that

$$\|u - v\| = \|A(w - y)\| = \lambda \|d\| = \lambda = \Phi(A, x, z).$$

■

Theorem 1 below gives a characterization of $\Phi(A)$ in terms of the minimum distance between a proper face F of $\mathrm{conv}(A)$ and its *complement polytope* $\mathrm{conv}(A \setminus F)$. This characterization motivates the name *facial distance* for the quantity $\Phi(A)$. The minimum distance expression for $\Phi(A)$ in Theorem 1 can be seen as a refinement of the so-called *vertex-facet distance* defined by Beck and Shtern [2]. More precisely if we specialize Beck and Shtern's construction of vertex-facet distance [2, Lemma 3.1] to our context and assume a suitable normalization for the facet defining hyperplanes for $\mathrm{conv}(A)$, then it follows that the vertex-facet distance of the polytope $\mathrm{conv}(A)$ is

$$\min_{F \in \mathrm{facets}(\mathrm{conv}(A))} \mathsf{dist}(\mathrm{affine}(F), \mathrm{vertices}(\mathrm{conv}(A \setminus F)))$$

provided that $A = \text{vertices}(\text{conv}(A))$.

As the statement of Theorem 1 shows, the quantity $\Phi(A)$ has a similar geometric expression. As a consequence of Theorem 1, it follows that when $A = \text{vertices}(\text{conv}(A))$ the facial distance $\Phi(A)$ is at least as large as the vertex-facet distance of $\text{conv}(A)$.

Theorem 1 *Assume* $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ *and at least two columns of* A *are different. Then*

$$\Phi(A) = \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ \emptyset \subsetneq F \subsetneq \text{conv}(A)}} \mathsf{dist}(F, \text{conv}(A \setminus F)).$$

Furthermore, if $F \in \text{faces}(\text{conv}(A))$ *minimizes the right hand side, then there exist* $x, z \in \Delta_{n-1}$ *such that* $Az \in F, Ax \in \text{conv}(A \setminus F)$ *and*

$$\Phi(A) = \Phi(A, x, z) = \max_{s \in S(x),\, a \in A} \langle p, s - a \rangle = \|A(x - z)\|$$

for $p = \frac{A(x-z)}{\|A(x-z)\|}$.

Proof: We first show that

$$\Phi(A) \geq \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ \emptyset \subsetneq F \subsetneq \text{conv}(A)}} \mathsf{dist}(F, \text{conv}(A \setminus F)). \tag{3}$$

To that end, assume $x, z \in \Delta_{n-1}$ are such that $A(x - z) \neq 0$ and let $d := \frac{A(x-z)}{\|A(x-z)\|}$. Let $p \in \mathbb{R}^m$ be a vector attaining the minimum in (1). Consider the face F of $\text{conv}(A)$ defined as

$$F = \operatorname*{Argmin}_{v \in \text{conv}(A)} \langle p, v \rangle .$$

Observe that $\emptyset \neq F \neq \text{conv}(A)$ because $\text{conv}(A)$ is a nonempty compact set and $Ax \notin F$. From Proposition 1 it follows that $\Phi(A, x, z) = \|u - v\|$ for some $v \in F$ and $u \in \text{conv}(A \setminus F)$. Therefore,

$$\mathsf{dist}(F, \text{conv}(A \setminus F)) \leq \|u - v\| = \Phi(A, x, z).$$

Since this holds for any $x, z \in \Delta_{n-1}$ such that $A(x - z) \neq 0$, inequality (3) follows.

Next we show the reverse inequality

$$\Phi(A) \leq \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ \emptyset \subsetneq F \subsetneq \text{conv}(A)}} \mathsf{dist}(F, \text{conv}(A \setminus F)). \tag{4}$$

To that end, assume F minimizes the right-hand-side in (4). Let $u \in \operatorname{conv}(A \setminus F)$ and $v \in F$ be such that

$$\mathsf{dist}(F, \operatorname{conv}(A \setminus F)) = \|u - v\|. \tag{5}$$

The optimality conditions for u in (5) imply that $\langle v - u, s - u\rangle \leq 0$ for all $s \in \operatorname{conv}(A \setminus F)$. Likewise the optimality conditions for v in (5) imply that $\langle u - v, t - v\rangle \leq 0$ for all $t \in F$. Let $G \in \operatorname{faces}(\operatorname{conv}(A \setminus F))$ be the face defined by

$$G := \{s \in \operatorname{conv}(A \setminus F) : \langle u - v, s - u\rangle = 0\}.$$

By taking a face of F if necessary, we can assume that

$$\langle u - v, t - v\rangle = 0 \text{ for all } t \in F.$$

Therefore,

$$\langle u - v, s - u\rangle = \langle u - v, t - v\rangle = 0 \ \text{ for all } \ t \in F, s \in G. \tag{6}$$

Next we claim that

$$\langle u - v, a - v\rangle \geq 0 \ \text{ for all } \ a \in A. \tag{7}$$

We prove this claim by contradiction. Assume $a \in A$ is such that $\langle u - v, a - v\rangle < 0$. Then $\langle u - v, a - u\rangle = \langle u - v, a - v\rangle - \|u - v\|^2 < 0$ and from (6) we get $a \notin F$. Hence for $\lambda > 0$ sufficiently small the point $u + \lambda(a - u) \in \operatorname{conv}(A \setminus F)$ satisfies

$$\|u + \lambda(a - u) - v\|^2 = \|u - v\|^2 + 2\lambda \langle u - v, a - u\rangle + \lambda^2 \|v\|^2 < \|u - v\|^2,$$

which contradicts (5). Thus (7) is proven.

Let $x, z \in \Delta_{n-1}$ be such that $u = Ax, v = Az$ and $S(x) = A \cap G$. The latter is possible since $u \in G$. We finish by observing that for $p = d = \frac{A(x-z)}{\|A(x-z)\|} = \frac{u-v}{\|u-v\|}$

$$\begin{aligned}
\Phi(A, x, z) &\leq \max_{s \in S(x),\, a \in A} \langle p, s - a\rangle \\
&\leq \max_{s \in G, a \in A} \langle p, s - a\rangle \\
&= \langle p, u - v\rangle \\
&= \|u - v\| \\
&= \mathsf{dist}(F, \operatorname{conv}(A \setminus F)).
\end{aligned}$$

The first step follows from the construction of $\Phi(A, x, z)$. The second step follows because $S(x) \subseteq G$. The third and fourth steps follow from (6) and (7) and the choice of p. The fifth step follows from (5). Finally, since $\Phi(A) \leq \Phi(A, x, z) \leq \mathsf{dist}(F, \mathrm{conv}(A \setminus F))$ inequality (4) follows. ■

From Theorem 1 we can readily compute the values of $\Phi(A)$ in the special cases detailed in the examples below. We use the following notation. Let $e \in \mathbb{R}^m$ denote the vector with all components equal to one, and for $i = 1, \ldots, m$ let $e_i \in \mathbb{R}^m$ denote the vector with i-th component equal to one and all others equal to zero.

Example 1 *Suppose $A = \{0, 1\}^m \subseteq \mathbb{R}^m$. By Theorem 1, induction, and symmetry it follows that*

$$\Phi(A) = \mathsf{dist}(0, \mathrm{conv}(A \setminus \{0\})) = \mathsf{dist}(0, \mathrm{conv}\{e_1, \ldots, e_m\}) = \frac{\|e\|}{m} = \frac{1}{\sqrt{m}}.$$

Example 2 *Let $A = \{e_1, \ldots, e_m\} \subseteq \mathbb{R}^m$. By Theorem 1 and the facial structure of* $\mathrm{conv}(A)$ *it follows that*

$$\begin{aligned}
\Phi(A) &= \min_{\emptyset \subsetneq S \subsetneq A} \mathsf{dist}(\mathrm{conv}(S), \mathrm{conv}(A \setminus S)) \\
&= \min_{\emptyset \subsetneq S \subsetneq A} \left\| \frac{\sum_{s \in S} s}{|S|} - \frac{\sum_{a \in A \setminus S} a}{|A \setminus S|} \right\| \\
&= \min_{r \in \{1, \ldots, m-1\}} \sqrt{\frac{m}{r(m-r)}} \\
&= \begin{cases} \frac{2}{\sqrt{m}} & \text{if } m \text{ is even} \\ \frac{2}{\sqrt{m - \frac{1}{m}}} & \text{if } m \text{ is odd.} \end{cases}
\end{aligned}$$

We note that the values for $\Phi(A)$ in the above examples match exactly the values of the *pyramidal width* defined by Lacoste-Julien and Jaggi [11]. Theorem 2 below shows that indeed the pyramidal width is identical to the condition measure $\Phi(A)$. As a byproduct of this identity, the original definition of pyramidal width given in [11] can be simplified.

Lacoste-Julien and Jaggi define the *pyramidal directional width* of a finite set $A \subseteq \mathbb{R}^m$ with respect to a direction $r \in \mathbb{R}^m \setminus \{0\}$ and a base point $u \in \mathrm{conv}(A)$ as

$$\mathrm{PdirW}(A, r, u) := \min_{S \in S_u} \max_{a \in A, s \in S} \left\langle \frac{r}{\|r\|}, a - s \right\rangle$$

where $S_u = \{S \subseteq A : u \in \text{conv}(S)\}$. Lacoste-Julien and Jaggi also define the *pyramidal width* of a set A as

$$\text{PWidth}(A) := \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ u \in F, r \in \text{cone}\,(F-u) \setminus \{0\}}} \text{PdirW}(F \cap A, r, u).$$

Observe that $r \in \text{cone}\,(F-u) \setminus \{0\}$ if and only if r is a positive multiple of some $v - u$ where $u, v \in F$ and $u - v \neq 0$. Therefore

$$\text{PWidth}(A) = \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ u, v \in F, u \neq v}} \text{PdirW}(F \cap A, v-u, u). \tag{8}$$

We note that in the original definition of the pyramidal directional width in [11] Lacoste-Julien and Jaggi use the following more restricted set $\tilde{S}_u$ instead of S_u

$$\tilde{S}_u = \{S \subseteq A : u \text{ is a proper convex combination of elements in } S\}.$$

The larger set S_u that we use above yields an equivalent definition of pyramidal directional width because for all $u \in \text{conv}(A)$ and $r \in \mathbb{R}^m \setminus \{0\}$

$$\min_{S \in S_u} \max_{a \in A, s \in S} \left\langle \frac{r}{\|r\|}, a - s \right\rangle = \min_{S \in \tilde{S}_u} \max_{a \in A, s \in S} \left\langle \frac{r}{\|r\|}, a - s \right\rangle.$$

Theorem 2 *Assume* $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ *and at least two columns of* A *are different. Then*

$$\Phi(A) = \min_{u, v \in \text{conv}(A), u \neq v} \text{PdirW}(A, v-u, u) = \text{PWidth}(A). \tag{9}$$

Proof: Assume $F \in \text{faces}(A)$ minimizes $\mathsf{dist}(F, \text{conv}(A \setminus F))$ and $x, z \in \Delta_{n-1}$ are chosen as in the second statement of Theorem 1 so that for $p = \frac{A(x-z)}{\|A(x-z)\|}$ we have

$$\Phi(A) = \Phi(A, x, z) = \max_{s \in S(x),\, a \in A} \langle p, s - a \rangle.$$

Therefore for $u = Ax$ and $v = Az$ we have $S(x) \in S_u$ and

$$\begin{aligned} \Phi(A) &= \max_{a \in A, s \in S(x)} \left\langle \frac{v-u}{\|v-u\|}, a - s \right\rangle \\ &\geq \min_{S \in S_u} \max_{a \in A, s \in S} \left\langle \frac{v-u}{\|v-u\|}, a - s \right\rangle \\ &= \text{PdirW}(A, v-u, u). \end{aligned}$$

Hence we conclude that $\Phi(A) \geq \min_{u,v\in\mathrm{conv}(A),u\neq v} \mathrm{PdirW}(A, v-u, u)$.

On the other hand, for $u, v \in \mathrm{conv}(A), u \neq v$ let $S \in S_u$ be such that

$$\mathrm{PdirW}(A, v-u, u) := \max_{a\in A, s\in S} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle.$$

Since $S \in S_u$, we have $u \in \mathrm{conv}(S)$ and thus there exists $x \in \Delta_{n-1}$ such that $S(x) \subseteq S$ and $Ax = u$. Let $z \in \Delta_{n-1}$ be such that $Az = v$. Taking $p = d = \frac{A(x-z)}{\|A(x-z)\|} = -\frac{v-u}{\|v-u\|}$ it follows that

$$\begin{aligned}
\Phi(A) \leq \Phi(A, x, z) &\leq \max_{s\in S(x),\, a\in A} \langle p, s-a\rangle \\
&= \max_{a\in A, s\in S(x)} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\
&\leq \max_{a\in A, s\in S} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\
&= \mathrm{PdirW}(A, v-u, u).
\end{aligned}$$

Consequently $\Phi(A) \leq \min_{u,v\in\mathrm{conv}(A),u\neq v} \mathrm{PdirW}(A, v-u, u)$ as well. Therefore the identity between the first two terms in (9) follows. To finish, observe that by (8)

$$\begin{aligned}
\mathrm{PWidth}(A) &= \min_{F\in\mathrm{faces}(\mathrm{conv}(A))} \min_{u,v\in\mathrm{conv}(F),u\neq v} \mathrm{PdirW}(F\cap A, v-u, u) \\
&= \min_{\substack{F\in\mathrm{faces}(\mathrm{conv}(A)) \\ \exists u,v\in F,\, u\neq v}} \Phi(F\cap A) \\
&= \Phi(A).
\end{aligned}$$

The second step follows by applying the identity between the first two terms in (9) to each $F \cap A$. The third step follows because on the one hand $\mathrm{conv}(A) \in \mathrm{faces}(\mathrm{conv}(A))$ implies

$$\min_{\substack{F\in\mathrm{faces}(\mathrm{conv}(A)) \\ \exists u,v\in F,\, u\neq v}} \Phi(F\cap A) \leq \Phi(A);$$

and on the other hand Theorem 1 implies

$$\min_{\substack{F\in\mathrm{faces}(\mathrm{conv}(A)) \\ \exists u,v\in F,\, u\neq v}} \Phi(F\cap A) \geq \Phi(A).$$

■

As we will see in Section 3 below, the following *localized* variant of $\Phi(A)$, which could be quite a bit larger than $\Phi(A)$ plays a role in the linear convergence rate of the Frank-Wolfe algorithm. Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \subseteq \mathbb{R}^{m\times n}$ and at least two columns of A are different. For $z \in \Delta_{n-1}$ let

$$\Phi(A, z) := \min_{x\in\Delta_{n-1}: A(x-z)\neq 0} \Phi(A, x, z),$$

and for nonempty $Z \subseteq \Delta_{n-1}$ let

$$\Phi(A, Z) := \inf_{z\in Z} \Phi(A, z).$$

Theorem 3 gives a localized version of Theorem 1 for $\Phi(A, Z)$. Its proof relies on the following localized version of Proposition 1. We omit the proof of Proposition 2 as it is a straightforward modification of the proof of Proposition 1.

Proposition 2 *Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$ and $x, z \in \Delta_{n-1}$ are such that $A(x-z) \neq 0$ and let $d := \frac{A(x-z)}{\|A(x-z)\|}$. If $F \in \text{faces}(\text{conv}(A))$ contains Az then*

$$\min_{p:\langle p,d\rangle=1} \max_{s\in S(x), a\in F\cap A} \langle p, s-a\rangle =$$
$$\max\left\{\lambda > 0 : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x), Ay \in F,\ A(w-y) = \lambda d\right\}.$$

Furthermore, if p minimizes the left hand side and $u = Aw, v = Ay$ maximize the right hand side then $v \in \text{conv}(B)$ and $u \in \text{conv}(A \setminus B)$ where $B = \underset{a\in F\cap A}{\text{Argmin}} \langle p, a\rangle$, and

$$\min_{p:\langle p,d\rangle=1} \max_{s\in S(x), a\in F\cap A} \langle p, s-a\rangle = \|u-v\|.$$

Theorem 3 *Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$ and at least two columns of A are different. Then for nonempty $Z \subseteq \Delta_{n-1}$*

$$\Phi(A, Z) \geq \min_{\substack{G\in\text{faces}(F)\\ \emptyset\neq G\neq\text{conv}(A)}} \mathsf{dist}(G, \text{conv}(A\setminus G))$$

where F is the smallest face of $\text{conv}(A)$ containing $AZ = \{Az : z \in Z\}$.

Proof: This is a straightforward modification of the proof of Theorem 1. Observe that for $x \in \Delta_{n-1}, z \in Z$ with $A(x-z) \neq 0$ and $d = \frac{A(x-z)}{\|A(x-z)\|}$

$$\Phi(A, x, z) = \min_{p:\langle p,d\rangle=1} \max_{s\in S(x), a\in A} \langle p, s-a\rangle \geq \min_{p:\langle p,d\rangle=1} \max_{s\in S(x), a\in F\cap A} \langle p, s-a\rangle.$$

Let $p \in \mathbb{R}^m$ be a vector attaining the minimum in the above right hand side and consider the face G of F defined as

$$G = \operatorname*{Argmin}_{v \in F} \langle p, v \rangle .$$

Observe that $\emptyset \neq G \neq \operatorname{conv}(A)$ because F is a nonempty compact set and $Ax \notin G$. From Proposition 2 it follows that for some $u \in \operatorname{conv}(A \setminus G)$ and $v \in G$

$$\begin{aligned}
\Phi(A, x, z) &= \min_{p:\langle p,d\rangle=1} \max_{s \in S(x), a \in A} \langle p, s - a \rangle \\
&\geq \min_{p:\langle p,d\rangle=1} \max_{s \in S(x), a \in F \cap A} \langle p, s - a \rangle \\
&= \|u - v\| \\
&\geq \mathsf{dist}(G, \operatorname{conv}(A \setminus G)).
\end{aligned}$$

Since this holds for all $x \in \Delta_{n-1}, z \in Z$, we conclude that

$$\Phi(A, Z) \geq \min_{G \in \operatorname{faces}(F)} \mathsf{dist}(G, \operatorname{conv}(A \setminus G)).$$

■

It is evident from Theorem 1 and Theorem 3 that $\Phi(A, Z)$ can be arbitrarily larger than $\Phi(A)$. In particular, consider $A = \begin{bmatrix} M & 0 & 0 & \cdots & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \end{bmatrix} \in \mathbb{R}^{2 \times n}$ with $M \gg 1$, and $Z = \{e_1\} \subseteq \Delta_{n-1}$. For this A and Z we have

$$\Phi(A, Z) \geq M \gg \frac{1}{n} - \frac{1}{n-1} = \min_{\substack{F \in \operatorname{faces}(\operatorname{conv}(A)) \\ \emptyset \subsetneq F \subsetneq \operatorname{conv}(A)}} \mathsf{dist}(F, \operatorname{conv}(A \setminus F))$$

because in this case $\{a_1\} = \{Ae_1\} = AZ$ is the smallest face of $\operatorname{conv}(A)$ containing AZ and the minimum in the right hand side is attained at the face $F = \{a_n\}$ of $\operatorname{conv}(A)$.

3 Frank Wolfe algorithm with away steps

We next present a fairly generic linear convergence result for a version of the Frank Wolfe algorithm with away steps. We emphasize that the statement in Theorem 4 can be found in or inferred from results already shown in [2, 11]. The goal of this section is to highlight three central premises that yield the proof of linear convergence, namely a first premise in the form of

a *slope bound*, a second premise in the form of a *decrease condition*, and a third premise on the *search direction* selected by the algorithm. As we detail below, the third premise is naturally tied to the condition measures $\Phi(A)$ and $\Phi(A, Z)$ discussed in Section 2.

Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ and consider the problem

$$\min_{u \in \operatorname{conv}(A)} f(u) \tag{10}$$

where $f : \operatorname{conv}(A) \to \mathbb{R}$ is a differentiable convex function. Throughout this section we assume that at least two columns of A are different as otherwise problem (10) is trivial.

We will rely on the following notation related to problem (10) throughout the rest of the paper. Let f^*, U^*, Z^* be defined as

$$f^* := \min_{u \in \operatorname{conv}(A)} f(u),\ U^* := \operatorname*{Argmin}_{u \in \operatorname{conv}(A)} f(u),\ Z^* := \{z \in \Delta_{n-1} : Az \in U^*\}.$$

Algorithm 1 Frank-Wolfe Algorithm with Away Steps

Pick $x_0 \in \Delta_{n-1}$; put $u_0 := Ax_0$; $k := 0$
for $k = 0, 1, 2, \ldots$ **do**
  $j := \operatorname*{argmin}_{i=1,\ldots,n} \langle \nabla f(u_k), a_i \rangle$; $\ell := \operatorname*{argmax}_{i \in I(x_k)} \langle \nabla f(u_k), a_i \rangle$
  if $\langle \nabla f(u_k), a_j - u_k \rangle < \langle \nabla f(u_k), u_k - a_\ell \rangle$ or $|I(x_k)| = 1$ **then**
    $v := a_j - u_k$; $w := e_j - x_k$; $\gamma_{\max} := 1$ (regular step)
  else
    $v := u_k - a_\ell$; $w := x_k - e_\ell$; $\gamma_{\max} := \frac{\langle e_\ell, x_k \rangle}{1 - \langle e_\ell, x_k \rangle}$ (away step)
  end if
  choose $\gamma_k \in [0, \gamma_{\max}]$
  $x_{k+1} := x_k + \gamma_k w$; $u_{k+1} := u_k + \gamma_k v = Ax_{k+1}$
end for

Theorem 4 gives a fairly generic linear convergence result for Algorithm 1. This result hinges on the following three premises.

Premise 1: *There exists $\mu > 0$ such that the objective function f satisfies the following bound: For all $u \in \operatorname{conv}(A)$ and $u^* \in U^*$*

$$\langle \nabla f(u), u - u^* \rangle \geq \|u - u^*\| \sqrt{2\mu(f(u) - f^*)}.$$

Premise 1 readily holds if f is strongly convex with parameter μ. In this case the optimality of u^*, the strong convexity of f, and the arithmetic-geometric inequality imply

$$\langle \nabla f(u), u - u^* \rangle \geq f(y) - f^* + \frac{\mu}{2}\|u - u^*\|^2 \geq \|u - u^*\|\sqrt{2\mu(f(u) - f^*)}.$$

From the error bound of Beck and Shtern [2, Lemma 2.5], it follows that Premise 1 also holds in the more general case when $f(u) = h(Eu) + \langle b, u \rangle$ for some strongly convex function h.

Premise 2: *There exists $L > 0$ such that for all $\gamma \in [0, \gamma_{\max}]$*

$$f(u_k + \gamma v) \leq f(u_k) + \langle \nabla f(u_k), \gamma v \rangle + \frac{L\gamma^2\|v\|^2}{2},$$

and the steplength γ_k at each iteration of Algorithm 1 satisfies

$$\gamma_k = \operatorname*{argmin}_{\gamma \in [0, \gamma_{\max}]} \left\{ \langle \nabla f(u_k), \gamma v \rangle + \frac{L\gamma^2\|v\|^2}{2} \right\} = \min \left\{ -\frac{\langle \nabla f(u_k), v \rangle}{L\|v\|^2}, \gamma_{\max} \right\}.$$

Premise 2 holds if ∇f is Lipschitz with constant L.

Premise 3: *There exists $c > 0$ such that the search direction v selected in Step 5 or Step 7 of Algorithm 1 satisfies*

$$-\frac{\langle \nabla f(u_k), v \rangle}{\langle \nabla f(u_k), d \rangle} \geq c$$

for all $d = \frac{u_k - u^*}{\|u_k - u^*\|}$ such that $u^* \in U^*$ and $\langle \nabla f(u_k), d \rangle > 0$.

Premise 3 holds at each iteration of the Frank-Wolfe Algorithm with Away Steps for $c = \frac{\Phi(A)}{2}$ as well as for the sharper bound $c = \frac{\Phi(A, Z^*)}{2}$. To see this, observe that as noted by [11], the choice between regular or away steps ensures

$$-\langle \nabla f(u_k), v \rangle \geq \frac{1}{2} \langle \nabla f(u_k), a_\ell - a_j \rangle > 0.$$

Let $z \in Z^*$ be such that $Az = u^* \in U^*$. Take $d = \frac{A(x_k - z)}{\|A(x_k - z)\|} = \frac{u_k - u^*}{\|u_k - u^*\|}$ and $p = \frac{\nabla f(u_k)}{\langle \nabla f(u_k), d \rangle}$. Observe that $\langle p, d \rangle = 1$ and thus from the construction of $\Phi(A, x, z)$ we get

$$\frac{\langle \nabla f(u_k), a_\ell - a_j \rangle}{\langle \nabla f(u_k), d \rangle} = \max_{\ell \in I(x_k), a \in A} \langle p, a_\ell - a \rangle \geq \Phi(A, x_k, z) \geq \Phi(A, Z^*).$$

Hence

$$-\frac{\langle \nabla f(u_k), v\rangle}{\langle \nabla f(u_k), d\rangle} \geq \frac{\langle \nabla f(u_k), a_\ell - a_j\rangle}{2\,\langle \nabla f(u_k), d\rangle} \geq \frac{\Phi(A, Z^*)}{2} \geq \frac{\Phi(A)}{2}.$$

We note that a similar, though a bit weaker, bound for c is given in [2] in terms of the vertex-facet distance of $\mathrm{conv}(A)$. (See [2, Corollary 3.1].)

Theorem 4 *Assume $x_0 \in \Delta_{n-1}$ in Step 1 of Algorithm 1 is a vertex of Δ_{n-1}. If Premise 1, Premise 2, and Premise 3 hold then the sequence of points $\{u_k : k = 0, 1, \dots\}$ generated by Algorithm 1 satisfies*

$$f(u_k) - f^* \leq (1 - r)^{k/2}\,(f(u_0) - f^*) \tag{11}$$

for $r = \min\left\{\frac{\mu c^2}{L \cdot \mathsf{diam}(A)^2}, \frac{1}{2}\right\}$. In particular, if $\mu \leq L$ then (11) holds for the solution-independent rate

$$r = \frac{\mu}{L} \cdot \frac{\Phi(A)^2}{4\mathsf{diam}(A)^2}$$

as well as for the sharper, though solution-dependent rate

$$r = \frac{\mu}{L} \cdot \frac{\Phi(A, Z^*)^2}{4\mathsf{diam}(A)^2}.$$

Proof: This proof is an adaptation of the proofs in [2, 11, 14]. We consider the following three possible cases separately: $\gamma_k < \gamma_{\max}$, $\gamma_k = \gamma_{\max} \geq 1$, and $\gamma_k = \gamma_{\max} < 1$.

Case 1: $\gamma_k < \gamma_{\max}$. Premise 2 and the choice of v imply that

$$f(u_{k+1}) \leq f(u_k) - \frac{\langle \nabla f(u_k), v\rangle^2}{2L\|v\|^2}. \tag{12}$$

Premise 3 and Premise 1 in turn yield

$$\langle \nabla f(u_k), v\rangle^2 \geq \frac{c^2\,\langle \nabla f(u_k), u_k - u^*\rangle^2}{\|u_k - u^*\|^2} \geq 2\mu c^2 (f(u_k) - f^*).$$

Plugging the last inequality into (12) we get

$$\begin{aligned} f(u_{k+1}) - f^* &\leq \left(1 - \frac{\mu c^2}{L\|v\|^2}\right)(f(u_k) - f^*) \\ &\leq \left(1 - \frac{\mu c^2}{L \cdot \mathsf{diam}(A)^2}\right)(f(u_k) - f^*). \end{aligned}$$

Case 2: $\gamma_k = \gamma_{\max} \geq 1$. Premise 2 and the choice of v imply that

$$f(u_{k+1}) = f(u_k + \gamma_{\max} v) \leq f(u_k) + \frac{\gamma_{\max}}{2} \langle \nabla f(u_k), v \rangle \leq f(u_k) + \frac{1}{2} \langle \nabla f(u_k), v \rangle .$$

On the other hand, the choice of v and the convexity of f yield

$$\langle \nabla f(u_k), v \rangle \leq \min_{u \in \mathrm{conv}(A)} \langle \nabla f(u_k), u - u_k \rangle \leq \langle \nabla f(u_k), u^* - u_k \rangle \leq f^* - f(u_k).$$

Consequently, when $\gamma_k = \gamma_{\max} \geq 1$ we have

$$f(u_{k+1}) - f^* \leq f(u_k) - f^* + \frac{1}{2}(f^* - f(u_k)) = \frac{1}{2}(f(u_k) - f^*).$$

Case 3: $\gamma_k = \gamma_{\max} < 1$. Premise 2 and the choice of v imply that

$$f(u_{k+1}) = f(u_k + \gamma_{\max} v) \leq f(u_k) + \frac{\gamma_{\max}}{2} \langle \nabla f(u_k), v \rangle \leq f(u_k).$$

Hence $f(u_{k+1}) - f^* \leq f(u_k) - f^*$.

To finish, it suffices to show that after N iterations, Case 3 occurs at most $N/2$ times. To that end, we apply the following clever argument from [11]. Each time $\gamma_k = \gamma_{\max} < 1$ we have $|I(x_{k+1})| \leq |I(x_k)| - 1$. On the other hand, each time $\gamma_k < \gamma_{\max}$ we have $|I(x_{k+1})| \leq |I(x_k)| + 1$ and each time $\gamma_k = \gamma_{\max} \geq 1$ we have $|I(x_{k+1})| \leq |I(x_k)|$. Since $|I(x_0)| = 1$ and $|I(x_k)| \geq 1$ for all $x_k \in \Delta_{n-1}$, it follows that after N iterations there must have been at least as many iterations with $\gamma_k < \gamma_{\max}$ as there were with $\gamma_k = \gamma_{\max} < 1$. In particular, the number of iterations with $\gamma_k = \gamma_{\max} < 1$ is at most $N/2$. ■

As we noted at the end of Section 2, the quantity $\Phi(A, Z^*)$ in Theorem 4 can be arbitrarily better (larger) than $\Phi(A)$. Observe that the solution-independent bound is simply the most conservative solution-dependent one for all possible solution sets $Z^* \subseteq \Delta_{n-1}$.

The linear convergence rate (11) with $r = \frac{\mu}{L} \cdot \frac{\Phi(A)^2}{4\mathsf{diam}(A)^2}$ matches the one given in [11, Theorem 1] when f is strongly convex with Lipschitz gradient. Furthermore, the rate (11) in terms of Premises 1 through 3 is essentially equivalent to the one derived [2, Theorem 3.1]. A subtle difference is that the latter uses a weaker bound for Premise 3 in terms of the vertex-facet distance [2, Corollary 3.1].

4 Convex quadratic objective

We next establish a linear convergence result similar to Theorem 4 for the case when the objective function is of the form

$$f(u) = \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle,$$

for an $m \times m$ symmetric positive semidefinite matrix Q and $b \in \mathbb{R}^m$. Consider problem (10) and Algorithm 1 for this objective function. In this case the steplength γ_k in Step 9 of Algorithm 1 can be easily computed via the following exact line-search:

$$\begin{aligned}\gamma_k &:= \operatorname*{argmin}_{\gamma\in[0,\gamma_{\max}]} f(u_k + \gamma v)\\ &= \begin{cases} \min\left\{\gamma_{\max}, -\frac{\langle Qu_k+b, v\rangle}{\langle v, Qv\rangle}\right\} & \text{if } \langle v, Qv\rangle > 0\\ \gamma_{\max} & \text{if } \langle v, Qv\rangle = 0. \end{cases}\end{aligned} \tag{13}$$

In this case linear convergence can be obtained from Theorem 4 and the error bound [2, Lemma 2.5]. However, the more explicit rate of convergence stated in Theorem 5 below can be obtained via a refinement of the proof of Theorem 4.

Theorem 5 is similar in spirit to the main result of Beck and Teboulle [3]. However, the treatment in [3] is concerned with the regular Frank-Wolfe algorithm for minimizing a quadratic objective function of the form $\|Mu - b\|^2$ over a compact convex domain. Their main linear convergence result in [3], namely [3, Proposition 3.2], requires that the minimizer is in the relative interior of the domain and that its objective value is zero. By contrast, Theorem 5 below does not require any assumption about the location of the minimizer or its objective value and applies to Algorithm 1, that is, the variant of Frank-Wolfe algorithm with away steps.

The statement and proof of Theorem 5 are modifications of those of Theorem 4. The linear convergence rate in Theorem 5 is stated in terms of a variant $\bar{\Phi}_g$ of Φ. Similar to the proof of Theorem 4, the proof Theorem 5 follows by putting together variants of the three premises described in Section 3.

The construction of the variant $\bar{\Phi}_g$ of Φ relies on the objects $Z(g)$ and $\|\cdot\|_g$ defined next. Assume $\bar{A} \in \mathbb{R}^{(m+1)\times n}$. For $g \in \mathbb{R}^m$ define $Z(g) \subseteq \Delta_{n-1}$ as follows

$$Z(g) := \operatorname*{Argmin}_{z\in\Delta_{n-1}} \left\langle \begin{bmatrix} g \\ 1 \end{bmatrix}, \bar{A}z \right\rangle.$$

For $\bar{v} = \begin{bmatrix} v \\ v_{m+1} \end{bmatrix} \in \mathbb{R}^{m+1}$ let

$$\|\bar{v}\|_g = \sqrt{\left\| \begin{bmatrix} I_m & 0 \end{bmatrix} \bar{v} \right\|^2 + \left| \begin{bmatrix} g^{\mathrm{T}} & 1 \end{bmatrix} \bar{v} \right|} = \sqrt{\|v\|^2 + |g^{\mathrm{T}} v + v_{m+1}|}.$$

Observe that $\|\bar{v}\|_g > 0$ if $\bar{v} \neq 0$.

For $x \in \Delta_{n-1}$ and $z \in Z(g)$ with $\bar{A}(x-z) \neq 0$ let $\bar{d} := \frac{\bar{A}(x-z)}{\|\bar{A}(x-z)\|_g}$ and define

$$\bar{\Phi}_g\left(\bar{A}, x, z\right) = \min_{p:\left\langle p, \bar{d}\right\rangle = 1} \max\left\{\langle p, \bar{a}_\ell - \bar{a}_j\rangle : \ell \in I(x), j \in \{1, \ldots, n\}\right\} \tag{14}$$

and

$$\bar{\Phi}_g\left(\bar{A}\right) := \min_{\substack{x \in \Delta_{n-1}, z \in Z(g) \\ \bar{A}(x-z) \neq 0}} \bar{\Phi}_g\left(\bar{A}, x, z\right).$$

The construction of $\bar{\Phi}_g$ in (14) resembles that of Φ in (1). The key difference is the type of normalization used in $\bar{d}$ that discriminates between the first m and the last components of $\bar{A}(x-z)$ via g. As Proposition 3 below shows, $\bar{\Phi}_g$ can be bounded below in terms of Φ. In particular, Proposition 3 shows that $\bar{\Phi}_g(\bar{A}) > 0$ if at least two columns of $\bar{A}$ are different.

Theorem 5 *Assume the objective function $f(u)$ in problem* (10) *is $f(u) = \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle$ where Q is an $m \times m$ symmetric positive semidefinite matrix and $b \in \mathbb{R}^m$. Assume the matrix $\bar{A} = \begin{bmatrix} Q^{1/2}A \\ b^{\mathrm{T}}A \end{bmatrix}$ has at least two different columns as otherwise problem* (10) *is trivial.*

If $x_0 \in \Delta_{n-1}$ in Step 1 of Algorithm 1 is a vertex of Δ_{n-1} and the steplength γ_k in Step 9 of Algorithm 1 is computed as in (13) *then the convergence rate* (11) *holds with*

$$r = \min\left\{\frac{\bar{\Phi}_g\left(\bar{A}\right)^2}{8\mathsf{diam}(Q^{1/2}A)^2}, \frac{1}{2}\right\},$$

where $g = Q^{1/2}Az$ for $z \in Z^*$, which does not depend on the specific choice of $z \in Z^*$.

Proof: This is a modification of the proof of Theorem 4. Let $u^* = Az \in U^*$ for $z \in Z^*$ and let $u_k = Ax_k$ denote the k-th iterate generated by Algorithm 1. The proof is a consequence of $Z^* \subseteq Z(g)$ and the following three premises that hold provided $u_k \neq u^*$.

Premise 1':

$$\langle \nabla f(u_k), u_k - u^* \rangle \geq \|\bar{A}(x_k - z)\|_g \sqrt{f(u_k) - f^*}.$$

Premise 2': If $\gamma_k = -\frac{\langle \nabla f(u_k), v \rangle}{\langle v, Qv \rangle} = -\frac{\langle Qu_k + b, v \rangle}{\langle v, Qv \rangle} < \gamma_{\max}$ then

$$f(u_k) - f(u_{k+1}) = \frac{\langle \nabla f(u_k), v \rangle^2}{2 \langle v, Qv \rangle}.$$

Premise 3':

$$-\langle \nabla f(u_k), v \rangle \geq \frac{\max\limits_{\ell \in I(x_k), j \in \{1,\ldots,n\}} \langle \nabla f(u_k), a_\ell - a_j \rangle}{2} \geq \frac{\bar{\Phi}_g(\bar{A})}{2} \cdot \frac{\langle \nabla f(u_k), u_k - u^* \rangle}{\|\bar{A}(x_k - z)\|_g}.$$

Indeed, by putting together Premise 2', Premise 3', and Premise 1' it follows that when $\gamma_k = -\frac{\langle v, Qu_k + b \rangle}{\langle v, Qv \rangle} < \gamma_{\max}$

$$\begin{aligned} f(u_k) - f(u_{k+1}) &= \frac{\langle \nabla f(u_k), v \rangle^2}{2 \langle v, Qv \rangle} \\ &\geq \frac{\bar{\Phi}_g(\bar{A})^2}{8\mathsf{diam}(Q^{1/2}A)^2} \cdot \frac{\langle \nabla f(u_k), u_k - u^* \rangle^2}{\|\bar{A}(x_k - z)\|_g^2} \\ &\geq \frac{\bar{\Phi}_g(\bar{A})^2}{8\mathsf{diam}(Q^{1/2}A)^2} \cdot (f(u_k) - f^*), \end{aligned}$$

and consequently

$$f(u_{k+1}) - f^* \leq \left(1 - \frac{\bar{\Phi}_g\left(\bar{A}\right)^2}{8\mathsf{diam}(Q^{1/2}A)^2}\right) \cdot (f(u_k) - f^*).$$

The rest of the proof follows exactly as the last two paragraphs of the proof of Theorem 4.

We next show that $g = Q^{1/2}Az$ does not depend on the specific choice of $z \in Z^*$ and $Z^* \subseteq Z(g)$.

Let $z, z' \in Z^*$. From the optimality of $u^* = Az$ and $u' = Az'$ we get both $0 \leq \langle Qu^* + b, u' - u^* \rangle$ and $0 \leq \langle Qu' + b, u^* - u' \rangle$. Hence

$$0 \leq -\left\langle Q(u^* - u'), u^* - u' \right\rangle = -\|Q^{1/2}(u^* - u')\|^2 \Rightarrow Q^{1/2}u^* = Q^{1/2}u'.$$

That is, $Q^{1/2}Az = Q^{1/2}Az'$. In other words, $g = Q^{1/2}Az$ does not depend on the specific choice of $z \in Z^*$.

Next we show $Z^* \subseteq Z(g)$. To that end, let $z \in Z^*$ and $u^* = Az$. The optimality of u^* implies that for all $u = Ax \in \operatorname{conv}(A)$

$$\begin{aligned}
0 &\leq \langle \nabla f(u^*), u - u^* \rangle \\
&= \langle Qu^* + b, u - u^* \rangle \\
&= \left\langle Q^{1/2}Az, Q^{1/2}(Ax - Az) \right\rangle + \langle b, Ax - Az \rangle \\
&= \left\langle g, Q^{1/2}Ax \right\rangle + \langle b, Ax \rangle - \left(\left\langle g, Q^{1/2}Az \right\rangle + \langle b, Az \rangle \right).
\end{aligned}$$

Hence $z \in Z(g)$. Since this holds for all $z \in Z^*$ we have $Z^* \subseteq Z(g)$.

We next show each of the above three premises. Observe that

$$\langle \nabla f(u_k), u_k - u^* \rangle = f(u_k) - f^* + \frac{1}{2} \langle u_k - u^*, Q(u_k - u^*) \rangle \geq f(u_k) - f^*,$$

and

$$\begin{aligned}
\langle \nabla f(u_k), u_k - u^* \rangle &= \langle Q(u_k - u^*), u_k - u^* \rangle + \langle Qu^* + b, u_k - u^* \rangle \\
&= \langle QA(x_k - z), A(x_k - z) \rangle + \langle QAz, A(x_k - z) \rangle + \langle b, A(x_k - z) \rangle \\
&= \|Q^{1/2}A(x_k - z)\|^2 + \left\langle g, Q^{1/2}A(x_k - z) \right\rangle + b^{\mathrm{T}}A(x_k - z) \\
&= \left\| \bar{A}(x_k - z) \right\|_g^2.
\end{aligned}$$

Thus

$$\langle \nabla f(u_k), u_k - u^* \rangle \geq \left\| \bar{A}(x_k - z) \right\|_g \cdot \sqrt{f(u_k) - f^*}$$

and so Premise 1' follows.

Premise 2' is an immediate consequence of (13) and the form of f.

The first inequality in Premise 3' follows from the choice of v. To prove the second inequality in Premise 3', take $\bar{d} := \frac{\bar{A}(x_k - z)}{\|\bar{A}(x_k - z)\|_g}$ and

$$p := \frac{\|\bar{A}(x_k - z)\|_g}{\langle \nabla f(u_k), u_k - u^* \rangle} \cdot \begin{bmatrix} Q^{1/2}u_k \\ 1 \end{bmatrix} = \frac{\left\| \bar{A}(x_k - z) \right\|_g}{\left\langle \begin{bmatrix} Q^{1/2}u_k \\ 1 \end{bmatrix}, \bar{A}(x_k - z) \right\rangle} \cdot \begin{bmatrix} Q^{1/2}u_k \\ 1 \end{bmatrix}.$$

This choice of p guarantees that $\langle p, \bar{d} \rangle = 1$. Furthermore, observe that for $i, j \in \{1, \ldots, n\}$

$$\left\langle \begin{bmatrix} Q^{1/2}u_k \\ 1 \end{bmatrix}, \bar{a}_i - \bar{a}_j \right\rangle = \langle Qu_k, a_i - a_j \rangle + \langle b, a_i - a_j \rangle = \langle \nabla f(u_k), a_i - a_j \rangle.$$

Hence the above choice of p yields

$$\max_{\ell\in I(x_k),j\in\{1,\ldots,n\}} \langle p, \bar{a}_\ell - \bar{a}_j\rangle = \frac{\|\bar{A}(x_k - z)\|_g}{\langle \nabla f(u_k), u_k - u^*\rangle} \cdot \max_{\ell\in I(x_k),j\in\{1,\ldots,n\}} \langle \nabla f(u_k), a_\ell - a_j\rangle .$$

On the other hand, the construction of $\bar{\Phi}_g(\bar{A}, x, z)$ yields

$$\max_{\ell\in I(x_k),j\in\{1,\ldots,n\}} \langle p, \bar{a}_\ell - \bar{a}_j\rangle \geq \bar{\Phi}_g(\bar{A}, x_k, z) \geq \bar{\Phi}_g(\bar{A}).$$

Putting the above two together we get

$$\max_{\ell\in I(x_k),j\in\{1,\ldots,n\}} \langle \nabla f(u_k), a_\ell - a_j\rangle \geq \bar{\Phi}_g(\bar{A}) \cdot \frac{\|\bar{A}(x_k - z)\|_g}{\langle \nabla f(u_k), u_k - u^*\rangle}.$$

■

As Example 6 in Section 6 shows, the minimum in the expression for r in Theorem 5 cannot be omitted because $\frac{\bar{\Phi}_g(\bar{A})^2}{8\mathsf{diam}(Q^{1/2}A)^2} > \frac{1}{2}$ can occur.

The next proposition establishes a property of $\bar{\Phi}_g$ similar to that stated in Proposition 1 for Φ. It also provides a bound on $\bar{\Phi}_g$ in terms of Φ. Proposition 3 relies on one more piece of notation. For $\bar{A} \in \mathbb{R}^{(m+1)\times n}$ and $g \in \mathbb{R}^m$ define

$$\delta(g) = \max_{x\in\Delta_{n-1}, z\in Z(g)} \left\langle \begin{bmatrix} g \\ 1 \end{bmatrix}, \bar{A}(x - z) \right\rangle = \max_{x,z\in\Delta_{n-1}} \left\langle \begin{bmatrix} g \\ 1 \end{bmatrix}, \bar{A}(x - z) \right\rangle .$$

Proposition 3 *Assume $\bar{A} \in \mathbb{R}^{(m+1)\times n}$ and $g \in \mathbb{R}^m$. Assume $x \in \Delta_{n-1}$ and $z \in Z(g)$ are such that $\bar{A}(x - z) \neq 0$, and let $\bar{d} := \frac{\bar{A}(x-z)}{\|\bar{A}(x-z)\|_g}$.*

(a) *The following identity holds*

$$\bar{\Phi}_g\left(\bar{A}, x, z\right) = \max\left\{\lambda : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ \bar{A}(w - y) = \lambda \bar{d}\right\}. \quad (15)$$

Furthermore, if p attains the minimum in (14) and $\bar{u} = \bar{A}w$, $\bar{v} = \bar{A}y$ maximize the right hand side in (15) then $\bar{v} \in \mathrm{conv}(\bar{B})$, for $\bar{B} = \mathrm{Argmin}_{\bar{a}\in\bar{A}} \langle p, \bar{a}\rangle$, and $\bar{u} \in \mathrm{conv}(\bar{A} \setminus \bar{B})$. Furthermore,

$$\bar{\Phi}_g(\bar{A}, x, z) = \|\bar{u} - \bar{v}\|_g.$$

(b) *If* $\delta(g) = 0$ *then* $\bar{\Phi}_g\left(\bar{A}, x, z\right) = \Phi(A, x, z)$ *for* $A = \begin{bmatrix} I_m & 0 \end{bmatrix} \bar{A}$.

Otherwise $\bar{\Phi}_g\left(\bar{A}, x, z\right) \geq \Phi(\hat{A}, x, z)$ *for* $\hat{A} = \begin{bmatrix} I_m & 0 \\ \frac{g^{\mathrm{T}}}{\sqrt{\delta(g)}} & \frac{1}{\sqrt{\delta(g)}} \end{bmatrix} \bar{A}$.

In particular, if at least two columns of $\bar{A}$ *are different then* $\bar{\Phi}_g\left(\bar{A}\right) > 0$ *for all* $g \in \mathbb{R}^m$.

Proof:

(a) The proof is a straightforward modification of the proof of Proposition 1 with $\bar{A}$ in place of A, $\bar{d}$ in place of d, and $\|\cdot\|_g$ in place of $\|\cdot\|$.

(b) Assume $w, y \in \Delta_{n-1}$ and $\lambda \in \mathbb{R}$. Observe that $\bar{A}(w - y) = \lambda d$ if and only if $\bar{A}(w - y) = t\bar{A}(x - z)$ for $t = \frac{\lambda}{\|\bar{A}(x-z)\|_g}$. Furthermore, for $r \neq 0$

$$\bar{A}(w - y) = t\bar{A}(x - z) \Leftrightarrow \begin{bmatrix} I_m & 0 \\ rg^{\mathrm{T}} & r \end{bmatrix} \bar{A}(w - y) = t \begin{bmatrix} I_m & 0 \\ rg^{\mathrm{T}} & r \end{bmatrix} \bar{A}(x - z). \tag{16}$$

Now consider two possible cases separately.

Case 1: $\delta(g) = 0$**.** In this case $\left\langle \begin{bmatrix} g \\ 1 \end{bmatrix}, \bar{A}y \right\rangle = \begin{bmatrix} g^{\mathrm{T}} & 1 \end{bmatrix} \bar{A}y$ is the same for all $y \in \Delta_{n-1}$ and consequently the last row of equations in (16) is redundant. Thus for $A = \begin{bmatrix} I_m & 0 \end{bmatrix} \bar{A}$

$$\bar{A}(w - y) = t\bar{A}(x - z) \Leftrightarrow A(w - y) = tA(x - z)$$

From part (a) and Proposition 1 we get

$$\begin{aligned} &\frac{\bar{\Phi}_g\left(\bar{A}, x, z\right)}{\|\bar{A}(x - z)\|_g} \\ &= \max\left\{t : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ \bar{A}(w - y) = t\bar{A}(x - z)\right\} \\ &= \max\left\{t : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ A(w - y) = tA(z - x)\right\} \\ &= \frac{\Phi(A, x, z)}{\|A(x - z)\|}. \end{aligned}$$

Furthermore, since $\delta(g) = 0$ we have $\|\bar{A}(x - z)\|_g = \|A(x - z)\|$ and consequently $\bar{\Phi}_g\left(\bar{A}, x, z\right) = \Phi(A, x, z)$.

Case 2: $\delta(g) > 0$**.** Let $\hat{A} := \begin{bmatrix} I_m & 0 \\ \frac{g^{\mathrm{T}}}{\sqrt{\delta(g)}} & \frac{1}{\sqrt{\delta(g)}} \end{bmatrix} \bar{A}$. Plugging $r = \frac{1}{\sqrt{\delta}}$ in (16) we get

$$\bar{A}(w-u) = t\bar{A}(x-z) \Leftrightarrow \hat{A}(w-u) = t\hat{A}(x-z).$$

From part (a) and Proposition 1 we get

$$\begin{aligned}
&\frac{\bar{\Phi}_g\left(\bar{A}, x, z\right)}{\|\bar{A}(x-z)\|_g} \\
&= \max\left\{t : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ \bar{A}(w-y) = t\bar{A}(x-z)\right\} \\
&= \max\left\{t : \exists w, y \in \Delta_{n-1},\ I(w) \subseteq I(x),\ \hat{A}(w-y) = t\hat{A}(x-z)\right\} \\
&= \frac{\Phi(\hat{A}, x, z)}{\|\hat{A}(x-z)\|}.
\end{aligned}$$

Since $0 \leq \begin{bmatrix} g^{\mathrm{T}} & 1 \end{bmatrix} \bar{A}(x-z) \leq \delta(g)$, it follows that

$$\begin{aligned}
\|\hat{A}(x-z)\|^2 &= \|A(x-z)\|^2 + \frac{\left(\begin{bmatrix} g^{\mathrm{T}} & 1 \end{bmatrix} \bar{A}(x-z)\right)^2}{\delta(g)} \\
&\leq \|A(x-z)\|^2 + |\begin{bmatrix} g^{\mathrm{T}} & 1 \end{bmatrix} \bar{A}(x-z)| \\
&= \|\bar{A}(x-z)\|_g^2
\end{aligned}$$

and so

$$\bar{\Phi}_g(\bar{A}, x, z) = \Phi(\hat{A}, x, z) \cdot \frac{\|\bar{A}(z-x)\|_g}{\|\hat{A}(x-z)\|} \geq \Phi(\hat{A}, x, z).$$

Finally, observe that if $\delta(g) = 0$ and two columns of $\bar{A}$ are different, then at least two columns of A as different as well. Thus

$$\bar{\Phi}_g(\bar{A}) \geq \Phi(A, Z(g)) \geq \Phi(A) > 0.$$

On the other hand, if $\delta(g) > 0$ and two columns of $\bar{A}$ are different, then at least two columns of $\hat{A}$ as different as well. Thus

$$\bar{\Phi}_g(\bar{A}) \geq \Phi(\hat{A}, Z(g)) \geq \Phi(\hat{A}) > 0.$$

In either case we have $\bar{\Phi}_g(\bar{A}) > 0$. ■

5 Composite convex objective

We next extend the main ideas from Section 4 to the case when the objective is a composite function of the form

$$f(u) = h(Eu) + \langle b, u \rangle ,$$

where $E \in \mathbb{R}^{p\times m}$, $b \in \mathbb{R}^m$, and $h : \mathbb{R}^p \to \mathbb{R}$ is a strongly convex function with Lipschitz gradient.

Consider problem (10) for this objective function. If L is an upper bound on the Lipschitz constant of ∇h, then

$$f(u_k + \gamma v) \leq f(u_k) + \gamma \langle \nabla f(u_k), v \rangle + \frac{L\gamma^2 \|Ev\|^2}{2}.$$

Consequently, γ_k in Step 9 of Algorithm 1 could be computed as follows

$$\begin{aligned} \gamma_k &:= \operatorname*{argmin}_{\gamma\in[0,\gamma_{\max}]} \left\{ \langle \gamma \nabla f(u_k), v \rangle + \frac{L\gamma^2\|Ev\|^2}{2} \right\} \\ &= \begin{cases} \min\left\{\gamma_{\max}, -\frac{\langle \nabla f(u_k), v\rangle}{L\|Ev\|^2}\right\} & \text{if } \ Ev \neq 0 \\ \gamma_{\max} & \text{if } \ Ev = 0. \end{cases} \end{aligned} \tag{17}$$

Once again, linear convergence can be obtained from Theorem 4 and the error bound [2, Lemma 2.5]. However, the more explicit rate of convergence in Theorem 6 holds.

Theorem 6 *Assume in problem* (10) *the objective is* $f(u) = h(Eu) + \langle b, u\rangle$ *where* h *is* μ*-strongly convex and* ∇h *is Lipschitz. Assume the matrix* $\bar{A} = \begin{bmatrix} EA \\ \frac{1}{\mu} b^{\mathrm{T}} A \end{bmatrix}$ *has at least two different columns as otherwise problem* (10) *is trivial.*

If $x_0 \in \Delta_{n-1}$ *in Step 1 of Algorithm 1 is a vertex of* Δ_{n-1} *and the steplength* γ_k *is computed as in* (17) *for some upper bound* L *on the Lipschitz constant of* ∇h *then the convergence rate* (11) *holds with*

$$r = \min\left\{ \frac{\mu}{L} \cdot \frac{\bar{\Phi}_g\left(\bar{A}\right)^2}{8\mathsf{diam}(EA)^2}, \frac{1}{2} \right\},$$

where $g = \frac{1}{\mu}\nabla h(EAz)$ *for* $z \in Z^*$*, which does not depend on the specific choice of* $z \in Z^*$.

Proof: This proof is a straightforward modification of the proof of Theorem 5. It is a consequence of $Z^* \subseteq Z(g)$ and the following three premises that hold provided $u_k \neq u^*$.

Premise 1":

$$\langle \nabla f(u_k), u_k - u^* \rangle \geq \|\bar{A}(x_k - z)\|_g \sqrt{\mu(f(u_k) - f^*)}.$$

Premise 2": If $\gamma_k = -\frac{\langle \nabla f(u_k), v \rangle}{L\|Ev\|^2} < \gamma_{\max}$ then

$$f(u_k) - f(u_{k+1}) \leq \frac{\langle \nabla f(u_k), v \rangle^2}{2L\|Ev\|^2}.$$

Premise 3":

$$-\langle \nabla f(u_k), v \rangle \geq \frac{\max\limits_{\ell \in I(x_k), j \in \{1,\ldots,n\}} \langle \nabla f(u_k), a_\ell - a_j \rangle}{2} \geq \frac{\bar{\Phi}_g(\bar{A})}{2} \frac{\langle \nabla f(u_k), u_k - u^* \rangle}{\|A(x_k - z)\|_g}.$$

We first show that $g = \frac{1}{\mu}\nabla h(EAz)$ does not depend on the specific choice of $z \in Z^*$ and $Z^* \subseteq Z(g)$.

Let $z, z' \in Z^*$. From the optimality of $u^* = Az$ and $u' = Az'$ we get both $0 \leq \langle \nabla f(u^*), u' - u^* \rangle$ and $0 \leq \langle \nabla f(u'), u^* - u' \rangle$. Since $\nabla f(y) = E^{\mathrm{T}} \nabla h(EAz) + b$ and h is μ-strongly convex we get

$$\begin{aligned} 0 &\leq -\left\langle \nabla f(u^*) - \nabla f(u'), u^* - u' \right\rangle \\ &= -\left\langle \nabla h(EAz) - \nabla h(EAz'), EA(z - z') \right\rangle \\ &\leq -\mu \|EA(z - z')\|^2. \end{aligned}$$

This implies that $EAz = EAz'$ and thus $g = \frac{1}{\mu}\nabla h(EAz)$ does not depend on the specific choice of $z \in Z^*$.

We next show $Z^* \subseteq Z(g)$. To that end, let $z \in Z^*$ and $u^* = Az$. The optimality of u^* implies that for all $y = Ax \in \operatorname{conv}(A)$

$$\begin{aligned} 0 &\leq \frac{1}{\mu} \langle \nabla f(u^*), u - u^* \rangle \\ &= \frac{1}{\mu} \left\langle E^{\mathrm{T}} \nabla h(Eu^*) + b, u - u^* \right\rangle \\ &= \langle g, EAx \rangle + \left\langle \frac{1}{\mu} b, Ax \right\rangle - \left(\langle g, EAz \rangle + \left\langle \frac{1}{\mu} b, Az \right\rangle \right). \end{aligned}$$

Hence $z \in Z(g)$. Since this holds for all $z \in Z^*$, we have $Z^* \subseteq Z(g)$.

We next show each of the above three premises. Since f is convex we readily have

$$\langle \nabla f(u_k), u_k - u^* \rangle \geq f(u_k) - f^*.$$

On the other hand, since h is μ-strongly convex we also have

$$\begin{aligned}
\langle \nabla f(u_k), u_k - u^* \rangle &= \langle \nabla f(u_k) - \nabla f(u^*), u_k - u^* \rangle + \langle \nabla f(u^*), u_k - u^* \rangle \\
&= \langle \nabla h(u_k) - \nabla h(u^*), E(u_k - u^*) \rangle + \langle \nabla f(u^*), u_k - u^* \rangle \\
&\geq \mu \|E(u_k - u^*)\|^2 + \langle \nabla h(Eu^*), E(u_k - u^*) \rangle + \langle b, u_k - u^* \rangle \\
&= \mu \left(\|EA(x_k - z)\|^2 + \langle g, EA(x_k - z) \rangle + \frac{1}{\mu} b^{\mathrm{T}} A(x_k - z) \right) \\
&= \mu \|\bar{A}(x_k - z)\|_g^2.
\end{aligned}$$

Therefore

$$\langle \nabla f(u_k), u_k - u^* \rangle \geq \|\bar{A}(x_k - z)\|_g \cdot \sqrt{\mu (f(u_k) - f^*)}$$

and Premise 1" follows.

Premise 2" is an immediate consequence of (17) and the fact the L is a bound on the Lipschitz constant of ∇h.

The proof of Premise 3" is also similar to that of Premise 3' in the proof of Theorem 5. The first inequality follows from the choice of v. For the second inequality take $\bar{d} := \frac{A(x_k - z)}{\|A(x_k - z)\|_g}$ and

$$p := \frac{\|\bar{A}(x_k - z)\|_g}{\langle \nabla f(u_k), u_k - u^* \rangle} \cdot \begin{bmatrix} \nabla h(Eu_k) \\ \mu \end{bmatrix} = \frac{\left\|\bar{A}(x_k - z)\right\|_g}{\left\langle \begin{bmatrix} \nabla h(Eu_k) \\ \mu \end{bmatrix}, \bar{A}(x_k - z) \right\rangle} \cdot \begin{bmatrix} \nabla h(Eu_k) \\ \mu \end{bmatrix}.$$

This choice of p guarantees that $\left\langle p, \bar{d} \right\rangle = 1$. Furthermore, observe that for all $i, j \in \{1, \ldots, n\}$

$$\left\langle \begin{bmatrix} \nabla h(Eu_k) \\ \mu \end{bmatrix}, \bar{a}_i - \bar{a}_j \right\rangle = \left\langle E^{\mathrm{T}} \nabla h(Eu_k), a_i - a_j \right\rangle + \langle b, a_i - a_j \rangle = \left\langle \nabla f(u_k), a_i - a_j \right\rangle.$$

Hence the above choice of p yields

$$\max_{\ell \in I(x_k), j \in \{1, \ldots, n\}} \langle p, \bar{a}_\ell - \bar{a}_j \rangle = \frac{\|\bar{A}(x_k - z)\|_g}{\langle \nabla f(u_k), u_k - u^* \rangle} \cdot \max_{\ell \in I(x_k), j \in \{1, \ldots, n\}} \left\langle \nabla f(u_k), a_\ell - a_j \right\rangle.$$

On the other hand, the construction of $\bar{\Phi}_g(\bar{A}, x, z)$ yields

$$\max_{\ell \in I(x_k), j \in \{1,\ldots,n\}} \langle p, \bar{a}_\ell - \bar{a}_j \rangle \geq \bar{\Phi}_g(\bar{A}, x_k, z) \geq \bar{\Phi}_g(\bar{A}).$$

Putting the above two together we get

$$\max_{\ell \in I(x_k), j \in \{1,\ldots,n\}} \langle \nabla f(u_k), a_\ell - a_j \rangle \geq \bar{\Phi}_g(\bar{A}) \cdot \frac{\langle \nabla f(u_k), u_k - u^* \rangle}{\|\bar{A}(x_k - z)\|_g}.$$

■

6 Some examples

Example 3 and Example 4 below show that the convergence bounds for the Frank-Wolfe algorithm with away steps stated in Theorem 4 and Theorem 5 in terms of Φ and $\bar{\Phi}_g$ are tight modulo some constants. Example 5 shows that the bound on $\bar{\Phi}_g$ in terms of Φ in Proposition 3 is also tight modulo a constant. Finally, Example 6 shows that the minimum in the expressions for r in Theorem 5 and Theorem 6 cannot be omitted.

We illustrate the bounds in Example 3 and Example 4 via computational experiments. Our experiments were conducted via a verbatim implementation of Algorithm 1 in matlab for a convex quadratic objective with steplength computed as in (13). The matlab code is publicly available at the following website `http://www.andrew.cmu.edu/user/jfp/fwa.html`. The reader can easily replicate the numerical results described below.

Example 3 *Let* $\theta \in (0, \pi/6)$ *and* $A := \begin{bmatrix} \cos(2\theta) & 1 & -1 \\ \sin(2\theta) & 0 & 0 \end{bmatrix}$. *In this case* $\mathsf{diam}(A) = 2$ *and* $\Phi(A) = \sin(\theta)$. *Consider the problem*

$$\min_{u \in \mathrm{conv}(A)} \frac{1}{2}\|u\|^2.$$

The optimal value of this problem is zero attained at $u^* = 0$. *Furthermore, the condition number of the objective function is one. If we apply Algorithm 1 to this problem starting with* $u_0 = a_1 = Ae_1$, *then it follows that for* $k = 1, 2, \ldots$ *the algorithm alternates between regular steps toward* a_2 *and away steps from* a_1. *Furthermore, it can be shown via a geometric reasoning that for* $k = 1, 2, \ldots$

$$\frac{1}{2}\|u_{k+1}\|^2 = \frac{1}{2}\|u_k\|^2 \cos^2(\theta_k)$$

where $\theta_k \in (0, 3\theta)$. *In particular, for* $k = 1, 2, \ldots$

$$\frac{\frac{1}{2}\|u_{k+1}\|^2}{\frac{1}{2}\|u_k\|^2} = 1 - \sin^2(\theta_k) \geq 1 - 9\sin^2(\theta) = 1 - \frac{36\Phi(A)^2}{\mathsf{diam}(A)^2}.$$

Thus the rate of convergence (11) *in Theorem 4 is tight modulo a constant.*

Figure 1 shows the ratio $1 - \frac{\frac{1}{2}\|u_{k+1}\|^2}{\frac{1}{2}\|u_k\|^2}$ *and the bound* $9\sin^2(\theta)$ *based on numerical experiments for* $\theta = \pi/10$, $\theta = \pi/100$, *and* $\theta = \pi/1000$. *The figure confirms that the above ratio stays bounded away from zero, that is, the objective values* $\frac{1}{2}\|u_k\|^2$ *converge linearly to zero. The figure also confirms that in each case the above ratio stays below and pretty close to the bound* $9\sin^2(\theta)$ *and thus the rate of linear convergence of* $\frac{1}{2}\|u_k\|^2$ *to zero is slower than* $1 - 9\sin^2(\theta)$.

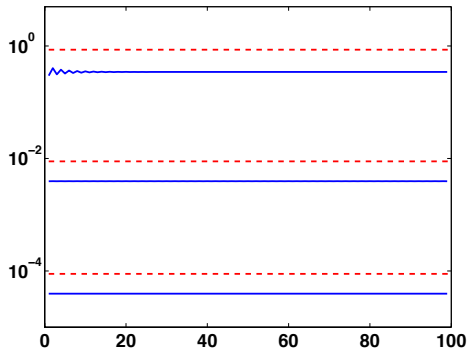

Figure 1: Plot of the ratio $1 - \frac{\frac{1}{2}\|u_{k+1}\|^2}{\frac{1}{2}\|u_k\|^2}$ (solid line) and the bound $9\sin^2(\theta)$ (dash line) in Example 3 for $\theta = \pi/10$ (top lines), $\theta = \pi/100$ (middle lines), and $\theta = \pi/1000$ (bottom lines).

Example 4 *Let* $t > 0$ *and* $A := \begin{bmatrix} t & t & -t \\ t & 0 & 0 \end{bmatrix}$. *Consider the problem*

$$\min_{u \in \mathrm{conv}(A)} \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle$$

where $Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ *and* $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. *The optimal value of this problem is zero attained at* $u^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. *Thus* $\bar{A} = \begin{bmatrix} Q^{1/2}A \\ b^{\mathrm{T}}A \end{bmatrix} = \begin{bmatrix} t & t & -t \\ 0 & 0 & 0 \\ t & 0 & 0 \end{bmatrix}$, $Z^* =$

$\{0.5e_2 + 0.5e_3\}$, *and* $g = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. *Proposition 3(a) and some straightforward calculations show that* $Z(g) = \operatorname{conv}\{e_2, e_3\}$, $\mathsf{diam}(Q^{1/2}A) = 2t$, *and*

$$\bar{\Phi}_g(\bar{A}) = \begin{cases} 2t & \text{if } t < 1/8 \\ \sqrt{t - 1/16} & \text{if } t \geq 1/8. \end{cases}$$

If we apply Algorithm 1 to this problem starting with $u_0 = a_1 = Ae_1$, *then it follows that for* $k = 1, 2, \ldots$ *the algorithm alternates between regular steps toward* a_2 *and away steps from* a_1. *Furthermore, for* $t \gg 1$ *it can be shown via a geometric reasoning that for* $1 \leq k < t/4$

$$\frac{\frac{1}{2}\langle u_{k+1}, Qu_{k+1}\rangle + \langle b, u_{k+1}\rangle}{\frac{1}{2}\langle u_k, Qu_k\rangle + \langle b, u_k\rangle} \geq 1 - \frac{4}{t}.$$

Observe that for $t \gg 1$ *we have*

$$\frac{\bar{\Phi}_g(\bar{A})^2}{8\mathsf{diam}(Q^{1/2}A)^2} = \frac{t - 1/16}{32t^2} \approx \frac{1}{32t}.$$

Therefore, the rate of convergence in Theorem 5 is tight modulo a constant. Notice that in sharp contrast to $\frac{\bar{\Phi}_g(\bar{A})}{\mathsf{diam}(Q^{1/2}A)}$ *which tends to zero as* $t \to \infty$, *all of* $\frac{\Phi(A)}{\mathsf{diam}(A)}$, $\frac{\Phi(Q^{1/2}A)}{\mathsf{diam}(Q^{1/2}A)}$, *and* $\frac{\Phi(\bar{A})}{\mathsf{diam}(Q^{1/2}A)}$ *stay constant and bounded away from zero for all* $t > 0$. *Thus the convergence rate in Theorem 5 cannot be stated solely in terms of any of the latter three ratios.*

Figure 2 shows the ratio $1 - \frac{\frac{1}{2}\langle u_{k+1}, Qu_{k+1}\rangle + \langle b, u_{k+1}\rangle}{\frac{1}{2}\langle u_k, Qu_k\rangle + \langle b, u_k\rangle}$ *and the bound* $\frac{4}{t}$ *based on numerical experiments for* $t = 200$, $t = 20000$, *and* $t = 2000000$. *Once again, the figure confirms that the objective values converge linearly to zero and that the ratio stays below and pretty close to the bound* $\frac{4}{t}$. *However, we should note that the latter only holds for* k *up to a certain threshold. Given the simplicity of this example, the optimal value is attained for* k *sufficiently large.*

Example 5 *Let* $t > 0$ *and* $\bar{A} := \begin{bmatrix} t & t & -t \\ t & 0 & 0 \end{bmatrix}$. *For* $g = 0$ *proceeding as in Example 4 it follows that* $Z(g) = \operatorname{conv}\{e_2, e_3\}, \delta(g) = t$, *and*

$$\bar{\Phi}_g(\bar{A}) = \begin{cases} 2t & \text{if } t < 1/8 \\ \sqrt{t - 1/16} & \text{if } t \geq 1/8. \end{cases}$$

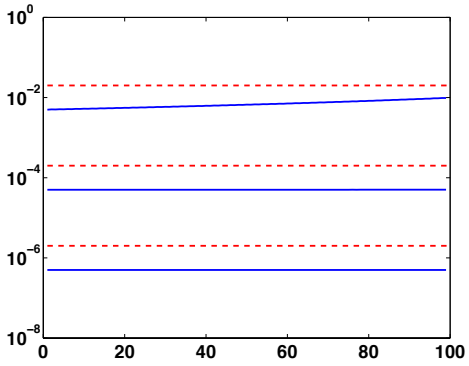

Figure 2: Plot of the ratio $1 - \frac{\frac{1}{2}\langle u_{k+1}, Qu_{k+1}\rangle + \langle b, u_{k+1}\rangle}{\frac{1}{2}\langle u_k, Qu_k\rangle + \langle b, u_k\rangle}$ (solid line) and the bound $\frac{4}{t}$ (dash line) in Example 4 for $t = 200$ (top lines), $t = 20000$ (middle lines), and $t = 2000000$ (bottom lines).

In this case the matrix $\hat{A}$ in Proposition 3(b) is

$$\hat{A} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{t}} \end{bmatrix} \bar{A} = \begin{bmatrix} t & t & -t \\ \sqrt{t} & 0 & 0 \end{bmatrix}.$$

Proposition 1 and some straightforward calculations show that

$$\Phi(\hat{A}, Z(g)) = \frac{2t}{\sqrt{4t+1}}.$$

This shows that the bound in Proposition 3(b) is tight modulo a constant. Indeed, notice that in this example $\frac{\Phi(\hat{A}, Z(g))}{\Phi_g(\bar{A})} \to 1$ both when $t \to \infty$ and when $t \downarrow 0$.

Example 6 *Let $t > 0$ and $A := \begin{bmatrix} 0 & 1 \\ 0 & t \end{bmatrix}$. Consider the problem*

$$\min_{u \in \mathrm{conv}(A)} \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle$$

where $Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The optimal value of this problem is zero attained at $u^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Thus $\bar{A} = \begin{bmatrix} Q^{1/2}A \\ b^{\mathrm{T}}A \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & t \end{bmatrix}$, $Z^* = \{e_1\}$, and

$g = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. *Proposition 3(a) and some straightforward calculations show that* $Z(g) = \text{conv}\{e_1\}$, $\mathsf{diam}(Q^{1/2}A) = 1$, *and* $\bar{\Phi}_g(\bar{A}) = \sqrt{1+t}$. *Thus for* $t \gg 1$

$$\frac{\bar{\Phi}_g\left(\bar{A}\right)^2}{8\mathsf{diam}(Q^{1/2}A)^2} = \frac{1+t}{8} > \frac{1}{2}.$$

Acknowledgements

Javier Peña's research has been supported by NSF grant CMMI-1534850.

References

[1] S. Ahipasaoglu, P. Sun, and M. Todd. Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.

[2] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Math. Program., To Appear.*

[3] A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Meth. of Oper. Res.*, 59(2):235–247, 2004.

[4] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.

[5] M. Epelman and R. M. Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Math. Program.*, 88(3):451–485, 2000.

[6] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Quarterly*, 3:95–110, 1956.

[7] D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *SIAM J. on Optim.*, 26:1493–1528, 2016.

[8] J. Guélat and P. Marcotte. Some comments on Wolfe's away step. *Math. Program.*, 35:110–119, 1986.

[9] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, volume 28 of *JMLR Proceedings*, pages 427–435, 2013.

[10] P. Kumar and E. A. Yildirim. A linearly convergent linear-time first-order algorithm for support vector classification with a core set result. *INFORMS Journal on Computing*, 23(3):377–391, 2011.

[11] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[12] R. Ñanculef, E. Frandi, C. Sartori, and H. Allende. A novel Frank-Wolfe algorithm. Analysis and applications to large-scale SVM training. *Inf. Sci*, 285:66–99, 2014.

[13] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer Academic Publishers, 2004.

[14] J. Peña, D. Rodríguez, and N. Soheili. On the von Neumann and Frank-Wolfe algorithms with away steps. *SIAM J. on Optim.*, 26:499–512, 2016.

[15] P. Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*. North-Holland, Amsterdam, 1970.