The condition number of a function relative to a set

David H. Gutman* Javier F. Peña†

December 5, 2019

Abstract

The *condition number* of a differentiable convex function, namely the ratio of its smoothness to strong convexity constants, is closely tied to fundamental properties of the function. In particular, the condition number of a quadratic convex function is the square of the aspect ratio of a canonical ellipsoid associated to the function. Furthermore, the condition number of a function bounds the linear rate of convergence of the gradient descent algorithm for unconstrained convex minimization.

We propose a condition number of a differentiable convex function relative to a reference set and distance function pair. This relative condition number is defined as the ratio of a *relative smoothness* to a *relative strong convexity constants*. We show that the relative condition number extends the main properties of the traditional condition number both in terms of its geometric insight and in terms of its role in characterizing the linear convergence of first-order methods for constrained convex minimization.

When the reference set X is a cone or a polyhedron and the function f is of the form $f = g \circ A$, we provide characterizations of and bounds on the condition number of f relative to X in terms of the usual condition number of g and a suitable condition number of the pair (A, X).

1 Introduction

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function. The *condition number* of f is the ratio L_f/μ_f where L_f and μ_f are respectively the *smoothness* and *strong convexity* constants of the function f. See Definition 1 and equation (9) below. The condition number L_f/μ_f is closely tied to a number of fundamental properties of the function f. In the special case when f is a quadratic convex function the condition

*Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, USA, david.gutman@ttu.edu

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

number has the following geometric interpretation. Suppose $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ where $A \in \mathbb{R}^{n\times n}$ is non-singular. Then the condition number of f is

$$\frac{L_f}{\mu_f} = \|A^\mathsf{T} A\| \cdot \|(A^\mathsf{T} A)^{-1}\| = (\|A\| \cdot \|A^{-1}\|)^2. \tag{1}$$

The latter quantity is the square of the aspect ratio of the ellipsoid $A(\mathbb{B}^n) := \{Ax : x \in \mathbb{R}^n, \|x\|_2 \leq 1\}$ since $\|A\|$ and $1/\|A^{-1}\|$ are respectively the radii of the smallest ball that contains $A(\mathbb{B}^n)$ and the largest ball contained in $A(\mathbb{B}^n)$.

The condition number L_f/μ_f also bounds the linear convergence rate of the gradient descent algorithm for the unconstrained minimization problem

$$f^\star = \min_{x\in\mathbb{R}^m} f(x).$$

More precisely, for a suitable choice of step sizes the iterates $x_k,\ k = 0, 1, \ldots$ generated by the gradient descent algorithm satisfy

$$\|X^\star - x_k\|_2^2 \leq \left(1 - \frac{\mu_f}{L_f}\right)^k \|X^\star - x_0\|_2^2$$

and

$$f(x_k) - f^\star \leq \frac{L_f}{2}\left(1 - \frac{\mu_f}{L_f}\right)^k \|X^\star - x_0\|_2^2,$$

where $X^\star := \{x \in \mathbb{R}^n : f(x) = f^\star\}$ and $\|X^\star - x\|_2 = \inf_{y\in X^\star} \|y - x\|_2$. The articles [4, 8, 17, 21–24], among others, discuss the above type of linear convergence and a number of interesting related developments. In particular, Necoara, Nesterov and Glineur [22] establish linear convergence properties for a wide class of first-order methods under assumptions that are relaxations of strong convexity.

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function, $X \subseteq \mathrm{dom}(f)$ be a convex set, and $D : X \times X \to \mathbb{R}_+$ a distance-like function, that is, $D(y, x) \geq 0$ and $D(x, x) = 0$ for all $x, y \in X$. We propose a relative smoothness constant $L_{f,X,D}$ and a relative strong convexity constant $\mu_{f,X,D}$ of the function f relative to the pair (X, D). See Definition 2 and equation (8) below for details. We show that the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ extends the above properties of the traditional condition number L_f/μ_f both in terms of its geometric insight and in terms of its role in characterizing the linear convergence of first-order methods for the constrained convex minimization problem

$$f^\star = \min_{x\in X} f(x). \tag{2}$$

As Example 1 illustrates, the relative condition number depends on the combination of the constraint set X and the function f. In particular, Example 1 shows that the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be vastly different (both smaller or larger) than the usual condition number L_f/μ_f depending on how the shape of X fits

f. Example 1 also shows that $\mu_{f,X,D}$ can be strictly positive in cases when $\mu_f = 0$. Our main results highlight deeper connections between the relative constants and geometric features of the set X. In particular, when $f = g \circ A$ for some matrix $A \in \mathbb{R}^{m\times n}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, and X is conic or polyhedral, we provide characterizations of and bounds on $L_{f,X,D}$ and $\mu_{f,X,D}$ in terms of L_g and μ_g and some condition properties of the pair (A, X).

We show that the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ and some related quantities readily yield linear convergence rates for the mirror descent, Frank-Wolfe, and Frank-Wolfe with Away Steps algorithms for the constrained minimization problem (2). We should note that the latter type of linear convergence properties have been previously established in [2, 3, 13, 18, 20, 22, 24, 28, 32] under various kinds of assumptions. Our approach shows that all of these linear convergence results hinge on a similar type of relative conditioning. Our approach also reveals that several linear convergence results can be sharpened. We show that the linear convergence of the mirror descent algorithm (Proposition 6 and Proposition 7) holds for a sharper rate and under more general assumptions that than those in [20,32]. More precisely, Proposition 6 and Proposition 7 show that linear convergence holds under new conditions of relative quasi-strong convexity and relative functional growth that are typically weaker than the type of relative strong convexity assumed in [20, 32]. In contrast to the previous results in [3, 13], our linear convergence result for the Frank-Wolfe algorithm (Proposition 8) is stated in terms of an affine invariant relative condition measure defined via a natural *radial* distance function. Our approach based on the relative condition number yields a proof of linear convergence for the Frank-Wolfe algorithm with away steps that is significantly shorter, simpler, and at least as sharp as or sharper than the ones previously presented in [2,18,28]. Unlike previous approaches, our proof of linear convergence of the Frank-Wolfe algorithm with away steps (Proposition 9) highlights some similarities with the proof of linear convergence of the regular Frank-Wolfe algorithm (Proposition 8). Like the results presented in [18, Appendix C and D], the linear convergence of the Frank-Wolfe algorithm with away steps (Proposition 9) is stated in terms of an affine invariant relative condition measure.

The relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ are defined *globally.* In particular, they do not depend on any specific point in X. We consider several variants of relative strong convexity following the constructions of Necoara, Nesterov and Glineur [22]. In particular, we define a *relative quasi-strong convexity constant* $\mu^{\star}_{f,X,D}$ and a *relative functional growth constant* $\mu^{\sharp}_{f,X,D}$. See Definition 3 and equation (12). Unlike $\mu_{f,X,D}$, the constants $\mu^{\star}_{f,X,D}$ and $\mu^{\sharp}_{f,X,D}$ depend on the set of minimizers $X^{\star}$ of f on X. We show that relative quasi-strong convexity is a relaxation of relative strong convexity. We also show that under suitable assumptions relative functional growth is a relaxation of relative quasi-strong convexity. Not surprisingly, there are classes of non-strongly convex functions for which the constant $\mu^{\sharp}_{f,X,D}$ is positive while $\mu_{f,X,D}$ and $\mu^{\star}_{f,X,D}$ may not be. (See Theorem 4.)

Our work draws on and connects several seemingly unrelated threads of research

on first-order methods [1, 2, 18, 20, 22, 28, 32] and on condition measures for convex optimization [9–12, 19, 25, 27, 30, 31]. Our construction of $L_{f,X,D}$ and $\mu_{f,X,D}$ is inspired by and closely related to the work of Lu, Freund, and Nesterov [20] and of Bauschke, Bolte, and Teboulle [1,32]. Lu et al. [20] extend the concepts of smoothness and strong convexity constants by considering them *relative* to a *reference* function h, see [20, Definition 1.1 and 1.2]. Our construction is identical to theirs in the special case when the distance function is the Bregman distance function D_h associated to a reference function h and the function f is strictly convex. Bauschke, Bolte, and Teboulle [1] define a concept of *Lipschitz-like* condition that is equivalent to smoothness relative to a reference function. As we detail in Section 5, our relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ are also identical to the *curvature constant, away curvature constant* and *geometric strong convexity constant* proposed by Jaggi [16] and by Lacoste-Julien and Jaggi in [18, Appendix C] for properly chosen distance-like functions D. Our constructions of relative functional growth, and relative quasi strong convexity are natural extensions of analogous concepts proposed by Necoara, Nesterov, and Glineur [22] to unveil relaxations of strong convexity that ensure the linear convergence of first-order methods. Our relative functional growth concept is in the same spirit as that of the quadratic functional growth approach used by Beck and Shtern [2] to established the linear convergence of a conditional gradient algorithm with away steps for non-strongly convex functions.

In contrast to the approaches in [2, 18, 20, 22, 28], our construction of the relative condition constants applies to any pair (X, D) of reference set and distance function. Our main results (Section 3 and Section 4) reveal some interesting insights when D is bounded by a squared norm. We establish a close connection between our relative conditioning approach and the conditioning of linear conic systems pioneered by Renegar [30, 31] and further developed by a number of authors [6, 9–12, 19, 25–27]. We especially draw on ideas developed in the recent paper [26]. We note that consistent with our construction of the relative constants $L_{f,X,D}$, $\mu_{f,X,D}$, $\mu^{\star}_{f,X,D}$, $\mu^{\sharp}_{f,X,D}$, all of our results concerning them scale appropriately, that is, they scale by λ whenever the objective function f is replaced by $\tilde{f} = \lambda f$ for some constant $\lambda > 0$. In particular, the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ and all of our bounds on it are invariant under positive scaling of f.

The main sections of the paper are organized as follows. Section 2 presents our central construction, namely relative smoothness and relative strong convexity. This section also introduces relative quasi strong convexity and relative functional growth, both of which are variants of relative strong convexity. Section 3 and Section 4 present the main technical results of the paper. Section 3 develops several properties of the constants $L_{f,X,D}$ and $\mu_{f,X,D}$. More precisely, Proposition 2 gives an upper bound on $L_{f,X,D}$ when f is of the form $g \circ A$ for some $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$. Proposition 2(a) shows that the bound is tight. The more involved Theorem 1 and Theorem 2 give lower bounds on $\mu_{f,X,D}$ when f is of the form $g \circ A$ and X is a convex cone or a polyhedron. These bounds readily imply that for $f = g \circ A$ the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be bounded in terms of the product of the classical condition number L_g/μ_g and a condition number of the pair (A, X). See equation (21)

and equation (24). Corollary 1 and Corollary 2 show that the bounds in Theorem 1 and Theorem 2 are tight. Section 4 develops properties analogous to those in Section 3 but for the constants $\mu^\star_{f,X,D}$ and $\mu^\sharp_{f,X,D}$. Section 5 details linear convergence results for the mirror descent algorithm, Frank-Wolfe algorithm, and Frank-Wolfe algorithm with away steps for problem (2). In all cases the linear convergence properties are stated in terms of the relative constants $L_{f,X,D}$ and $\mu^\star_{f,X,D}, \mu^\sharp_{f,X,D}$ for suitable choices of distance-like function D. The main results in Section 5 can be summarized as follows. Consider the mirror descent algorithm for problem (2) with a Bregman distance D_h associated to a reference function $h : X \to \mathbb{R}$. Proposition 6 shows the following linear convergence result: if $L_{f,X,D_h} < \infty$ and $\mu^\star_{f,X,D_h} > 0$ then the mirror descent iterates satisfy

$$f(x_k) - f^\star \leq L_{f,X,D_h} \left(1 - \frac{\mu^\star_{f,X,D_h}}{L_{f,X,D_h}}\right)^k D_h(x^\star, x_0)$$

for $x^\star \in \operatorname{argmin}_{x \in X} f(x)$. Proposition 7 gives a linear convergence result of similar flavor when $\mu^\sharp_{f,X,D_h} > 0$. The rates of convergence in both Proposition 6 and Proposition 7 are at least as sharp, and possibly much sharper, than those in [20, 32] and apply to a broader class of functions. In particular, as Example 7 in Section 4 shows, there are instances where $\mu^\sharp_{f,X,D} > \mu_{f,X,D} = 0$ occurs. In such instances Proposition 7 yields the linear convergence of mirror descent whereas the linear convergence results in [20, 32] do not apply.

Proposition 8 gives a strikingly similar linear convergence result for the Frank-Wolfe algorithm: suppose X is a compact convex set endowed with a linear oracle and $L_{f,X,\mathfrak{R}} < \infty$ and $\mu^\star_{f,X,\mathfrak{R}} > 0$ for the *radial distance function* $\mathfrak{R} : X \times X \to \mathbb{R}_+$ defined via (46). Proposition 8 shows that the Frank-Wolfe iterates satisfy

$$f(x_k) - f^\star \leq \left(1 - \frac{\mu^\star_{f,X,\mathfrak{R}}}{L_{f,X,\mathfrak{R}}}\right)^k (f(x_0) - f^\star).$$

This rate is convergence subsumes and is sharper than the previously known linear convergence results for the Frank-Wolfe algorithm in [3, 13].

Proposition 9 gives a result of similar flavor for the Frank-Wolfe algorithm with away steps: suppose X is a polytope endowed with a vertex linear oracle, and $L_{f,X,\mathfrak{D}} < \infty$ and $\mu^\star_{f,X,\mathfrak{G}} > 0$ for some suitable distance functions $\mathfrak{D} : X \times X \to \mathbb{R}_+$ and $\mathfrak{G} : X \times X \to \mathbb{R}_+$ defined via (49) and (51). Proposition 9 shows that if the Frank-Wolfe algorithm with away steps starts from a vertex in X then the subsequent iterates satisfy

$$f(x_k) - f^\star \leq \left(1 - \min\left\{\frac{1}{2}, \frac{\mu^\star_{f,X,\mathfrak{G}}}{4L_{f,X,\mathfrak{D}}}\right\}\right)^{k/2} (f(x_0) - f^\star).$$

This rate of convergence is at least as sharp, and possible much sharper, than the rates previously shown in [2, 18, 28].

Throughout the paper we define a number of new objects that are necessary for our main developments. To help the reader recall the definition and notation associated to these new objects, Table 1 displays the section and equation where each object is defined.

Symbol	Section	Equation
$Z_{f,X}(y)$	2	(3)
$L_{f,X,D}$ and $\mu_{f,X,D}$	2.1	(8)
L_f and μ_f	2.1	(9)
$\mu^{\star}_{f,X,D}$ and $\mu^{\sharp}_{f,X,D}$	2.2	(12)
$Z_{A,X}(y)$	3	(13)
$A\|C$ and $(A\|C)^{-1}$	3	(14) and (15)
$\|A\|C\|$ and $\|(A\|C)^{-1}\|$	3	(16)
$\mathcal{T}(A\|X)$	3.2	(22)
$\Phi(A)$ and $\mathsf{diam}(A)$	3.2	(25) and (26)
$\mathcal{T}(A\|X,S)$	4.1	(32)

Table 1: Index of symbols introduced in the paper

2 Conditioning relative to a reference set and distance function pair

This section presents the central ideas of this paper. We introduce the concepts of relative smoothness and relative strong convexity of a function relative to a reference set and distance function pair. We also introduce some variants of relative strong convexity that are natural extensions of the approach developed by Necoara, Nesterov and Glineur [22].

Throughout the entire paper we will typically make the following blanket assumption about the triple (f, X, D).

Assumption 1. The function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable. The set $X \subseteq \mathrm{dom}(f)$ is convex. The function $D : X \times X \to \mathbb{R}_+$ is a reference *distance-like* function, that is, $D(y, x) \geq 0$ for all $x, y \in X$ and $D(x, x) = 0$ for all $x \in X$.

Throughout our developments we will consider the following classes of reference distance-like functions:

- The *Bregman distance* $D_h : X \times X \to \mathbb{R}_+$ associated to a reference convex differentiable function $h : X \to \mathbb{R}$, that is,
$$D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle .$$
- The square of a (non-necessarily Euclidean) norm $\|\cdot\|$ in $\mathbb{R}^n$, that is,
$$D(y, x) := \frac{1}{2}\|y - x\|^2.$$
- The square $\mathfrak{R} := \frac{\mathfrak{r}^2}{2}$ of the *radial* distance function $\mathfrak{r} : X \times X \to \mathbb{R}_+$ defined as follows
$$\mathfrak{r}(y, x) := \inf\{\rho > 0 : y - x = \rho \cdot (u - x) \text{ for some } u \in X\}.$$

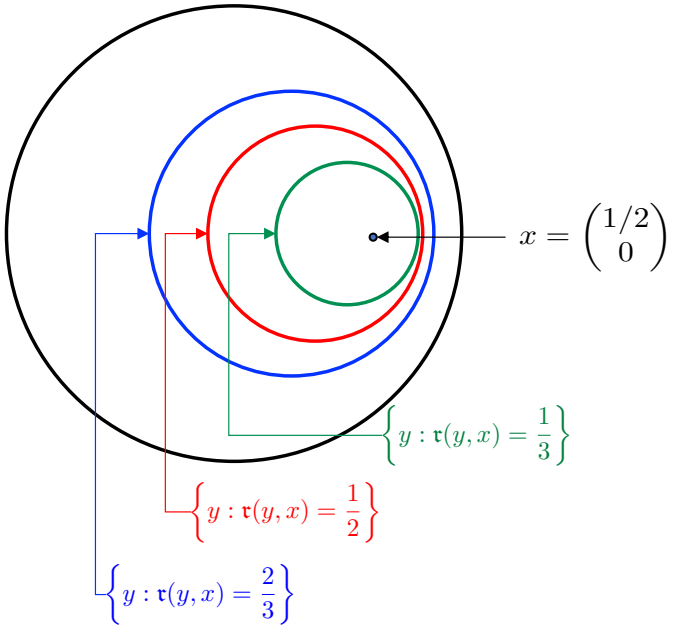

Figure 1: Level sets of $\mathfrak{r}(\cdot, x)$ in $X = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$.

Figure 1 illustrates the level sets defined by $\mathfrak{r}(\cdot, x)$ for $X = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$.

- The square $\mathfrak{D} := \frac{\mathfrak{d}^2}{2}$ of the *diametral* distance function $\mathfrak{d} : X \times X \to \mathbb{R}_+$ defined as follows

$$\mathfrak{d}(y, x) := \inf\{\delta > 0 : y - x = \delta \cdot (u - v) \text{ for some } u, v \in X\}.$$

Figure 2 illustrates the level sets defined by the diametral distance $\mathfrak{d}(\cdot, x)$ for $X = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$.

Our main construction is based on bounding the behavior of the Bregman distance associated to f in terms of the reference distance function D. The following set-valued mapping $Z_{f,X} : X \rightrightarrows X$ provides a key building block for our construction. For $y \in X$ let $Z_{f,X}(y) \subseteq X$ denote the set

$$Z_{f,X}(y) := \{x \in X : f(x) = f(y) \text{ and } \langle \nabla f(x) - \nabla f(y), x - y \rangle = 0\}. \tag{3}$$

It is easy to see that $Z_{f,X}(y)$ can also be written as

$$Z_{f,X}(y) = \{x \in X : f(x + \lambda(y - x)) = f(y) \text{ for all } \lambda \in [0, 1]\}.$$

Observe that if f is strictly convex then $Z_{f,X}(y) = \{y\}$ for all $y \in X$. The set $Z_{f,X}(y)$ captures the largest convex subset of $\{x \in X : f(x) = f(y)\}$ that includes y and where f fails to be strictly convex. In particular, when f is of the form $f = g \circ A$ for

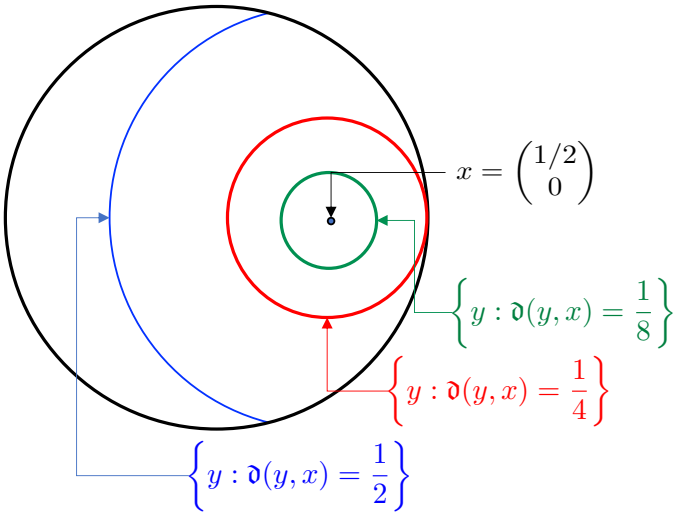

Figure 2: Level sets of $\mathfrak{d}(\cdot, x)$ in $X = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$.

$A \in \mathbb{R}^{m \times n}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ strictly convex, it is easy to see that $Z_{f,X}(y) = \{x \in X : Ax = Ay\}$. We will further discuss functions of this form in Section 3 and Section 4. To illustrate the set-valued mapping $Z_{f,X}$ in a different example, consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(x) := \min_{y \in \mathbb{B}^n} \|x - y\|_2^2,$$

where $\mathbb{B}^n = \{y \in \mathbb{R}^n : \|y\|_2 \leq 1\}$. In this case

$$Z_{f,X}(y) = \begin{cases} \{y\} & \text{if } y \notin \mathbb{B}^n \\ \mathbb{B}^n & \text{if } y \in \mathbb{B}^n. \end{cases}$$

2.1 Relative smoothness and relative strong convexity

To motivate our main construction we first recall the classical notion of smoothness and strong convexity constants. We recall these classical concepts in a format that we subsequently use for our main construction. Recall that for a convex differentiable function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $x, y \in \operatorname{dom}(f)$ the Bregman distance $D_f(y, x)$ is

$$D_f(y, x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle .$$

Definition 1. Suppose (f, X, D) satisfy Assumption 1 and $D(y, x) = \frac{1}{2}\|y - x\|^2$ for some norm $\|\cdot\|$ in $\mathbb{R}^n$.

(a) The function f is smooth on X if there exists a constant $L > 0$ such that

$$D_f(y, x) \leq L D(y, x) \quad \text{for all} \quad x, y \in X. \tag{4}$$

(b) The function f is strongly convex on X if there exists a constant $\mu > 0$ such that

$$D_f(y, x) \geq \mu D(y, x) \text{ for all } x, y \in X. \tag{5}$$

Next, we present our main construction. In Definition 2 and throughout the paper we will use the following notational convention. For a nonempty $S \subseteq X$ and $x \in X$ let $D_f(S, x)$ and $D(S, x)$ denote $\inf_{y \in S} D_f(y, x)$ and $\inf_{y \in S} D(y, x)$ respectively.

Definition 2. Let (f, X, D) satisfy Assumption 1.

(a) We say that f is *smooth relative* to (X, D) if there exists a constant $L > 0$ such that

$$D_f(y, x) \leq L D(y, x) \text{ for all } x, y \in X. \tag{6}$$

(b) We say that f is *strongly convex relative* to (X, D) if there exists a constant $\mu > 0$ such that

$$D_f(Z_{f,X}(y), x) \geq \mu D(Z_{f,X}(y), x) \text{ for all } x, y \in X. \tag{7}$$

When $D = D_h$ for some convex differentiable function $h : X \to \mathbb{R}$, the above relative smoothness concept is identical to the smoothness of f relative to h on X as defined in [20]. The latter in turn is equivalent to the *Lipschitz-like condition* defined in [1]. Furthermore, when $D = D_h$ and f is strictly convex, the above relative strong convexity concept is identical to the strong convexity of f relative to h on X as defined in [20]. We note that as in [20], the above definitions (6) and (7) are not symmetric in x and y since they depend on D_h and D which are not necessarily symmetric. Observe that the term $Z_{f,X}(y)$ instead of y in (7) makes this definition of relative strong convexity less stringent than the classical one (5) or the one in [20]. This is a key feature of our construction.

We will use the following notation throughout the rest of the paper. Suppose (f, X, D) satisfies Assumption 1. Let $L_{f,X,D}$ and $\mu_{f,X,D}$ be the following relative smoothness and strong convexity constants

$$L_{f,X,D} := \inf\{L > 0 : (6) \text{ holds}\}, \ \mu_{f,X,D} := \sup\{\mu > 0 : (7) \text{ holds}\}. \tag{8}$$

In addition, suppose (f, X, D) satisfies Assumption 1 and $D(x, y) = \frac{1}{2}\|x - y\|^2$ for some norm $\|\cdot\|$ in $\mathbb{R}^n$. Let L_f and μ_f be the following classical smoothness and strong convexity constants

$$L_f := \inf\{L > 0 : (4) \text{ holds}\}, \ \mu_f := \sup\{\mu > 0 : (5) \text{ holds}\}. \tag{9}$$

The following example illustrates the values of the relative smoothness and convexity constants $L_{f,X,D}$ and $\mu_{f,X,D}$ of a convex quadratic function relative to (X, D) for some canonical choices of f, X, and D. Example 1 highlights that the relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ depend on the combination of the constraint set X and the function f. In particular, Example 1 shows that the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be vastly different (both smaller or larger) than the usual condition number L_f/μ_f depending on how the shape of X fits f. Example 1 also lays the ground for the main properties that we develop in Section 3.

Example 1. Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ with $A \neq 0$ and $\mathbb{R}^n$ and $\mathbb{R}^n$ be endowed with the Euclidean norm. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $D(y, x) := \frac{1}{2}\|y - x\|_2^2$. Then f has the following smoothness and strong convexity constants $L_{f,X,D}$ and $\mu_{f,X,D}$ relative to (X, D) for some particular choices of X.

(a) For $X = \mathbb{R}^n$ we have $L_{f,X,D} = \sigma_{\max}(A^\mathsf{T}A) = \sigma_{\max}(A)^2$ and $\mu_{f,X,D} = \sigma^+_{\min}(A^\mathsf{T}A) = \sigma^+_{\min}(A)^2 > 0$, where $\sigma^+_{\min}(\cdot)$ denotes the smallest *positive* singular value. Observe that in this case $L_f = L_{f,X,D}$ but $\mu_f = \mu_{f,X,D}$ only when A is full column rank.

(b) Suppose $X \subseteq \mathbb{R}^n$ is a linear subspace such that the mapping $A|X : X \to \mathbb{R}^m$ defined via $x \in X \mapsto Ax \in \mathbb{R}^m$ is nonzero. Then $L_{f,X,D} = \sigma_{\max}(A|X)^2$ and $\mu_{f,X,D} = \sigma^+_{\min}(A|X)^2$. Observe that in this case $L_{f,X,D} \leq L_f$ and $L_{f,X,D}$ can be quite a bit smaller. Likewise, $\mu_{f,X,D} \geq \mu_f$ and $\mu_{f,X,D}$ could be quite a bit larger.

For instance, suppose $A = \mathrm{diag}(I_{n-2}, M, \epsilon) \in \mathbb{R}^{n\times n}$ for some positive M, ϵ with $0 < \epsilon \ll 1 \ll M$. If $X = \mathbb{R}^{n-2} \times \{0_2\} \subseteq \mathbb{R}^n$ then

$$\mu_f = \epsilon^2 \ll 1 = \mu_{f,X,D} = L_{f,X,D} \ll M^2 = L_f.$$

In this case we have $L_{f,X,D}/\mu_{f,X,D} \ll L_f/\mu_f$.

(c) Suppose $X = \mathbb{R}^n_+$. In this case $L_{f,X,D} = \|A\|^2 = \sigma_{\max}(A^\mathsf{T}A) = L_f$. On the other hand, if $A(\mathbb{R}^n_+) = \mathbb{R}^m$ then $\mu_{f,X,D}$ is the following kind of *signed* smallest singular value of A

$$\mu_{f,X,D} = \max\{r : r\mathbb{B}^m \subseteq A(\mathbb{B}^n \cap \mathbb{R}^n_+)\},$$

where $\mathbb{B}^m$ and $\mathbb{B}^n$ denote the unit balls in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. In other words, $\mu_{f,X,D}$ is the radius of the largest ball centered at zero and contained in $A(\mathbb{B}^n \cap \mathbb{R}^n_+)$. Observe that if $X = \mathbb{R}^n_+$ and $A(\mathbb{R}^n_+) = \mathbb{R}^m$ then $0 < \mu_{f,X,D} \leq \sigma_{\min}(A)$ and $\mu_{f,X,D}$ can be quite a bit smaller. For instance, if $A = \begin{bmatrix} 1 & -1 & 0 \\ -\epsilon & -\epsilon & 1 \end{bmatrix}$ for $0 < \epsilon \ll 1$ then

$$\mu_{f,X,D} = 2\epsilon^2 \ll 1 + \epsilon^2 = \sigma_{\min}(A) = \mu_f.$$

In this case we have $L_{f,X,D}/\mu_{f,X,D} \gg L_f/\mu_f$.

The statements (a), (b), and (c) in Example 1 can be verified directly but they also follow from the more general Proposition 2, Corollary 1, and Corollary 2 in Section 3 below.

2.2 Relative quasi strong convexity and relative functional growth

Following [22], we next consider two variants of relative strong convexity that are natural extensions of the *quasi-strong convexity* and *quadratic functional growth* concepts defined in [22]. For that purpose, we will rely on the following strengthening of Assumption 1.

Assumption 2. Suppose (f, X, D) satisfy Assumption 1, $f^\star := \min_{x\in X} f(x)$ is finite, $X^\star := \{x \in X : f(x) = f^\star\} \neq \emptyset$, and the map $x \mapsto \bar{x} := \operatorname{argmin}_{y\in X^\star} D(y, x)$ is well defined for all $x \in X$.

Definition 3. Suppose (f, X, D) satisfies Assumption 2.

(a) We say that f is *quasi-strongly-convex relative* to (X, D) if there exists a constant $\mu > 0$ such that
$$D_f(\bar{x}, x) \geq \mu D(\bar{x}, x) \text{ for all } x \in X. \tag{10}$$

(b) We say that f has *D-relative functional growth* on X if there exists a constant $\mu > 0$ such that
$$f(x) - f^\star \geq \mu D(\bar{x}, x) \text{ for all } x \in X. \tag{11}$$

Throughout the sequel we will use the following notation analogous to (8). Suppose (f, X, D) satisfies Assumption 2. Let $\mu^\star_{f,X,D}$ and $\mu^\sharp_{f,X,D}$ be as follows
$$\mu^\star_{f,X,D} := \sup\{\mu > 0 : (10) \text{ holds}\},\ \mu^\sharp_{f,X,D} := \sup\{\mu > 0 : (11) \text{ holds}\}. \tag{12}$$

The next proposition shows that, as one may intuitively expect, relative quasi-strong convexity is a relaxation of relative strong convexity. In other words, $\mu_{f,X,D} \leq \mu^\star_{f,X,D}$ whenever (f, X, D) satisfies Assumption 2.

Proposition 1. *Suppose (f, X, D) satisfy Assumption 2. If $\mu > 0$ is such that (f, X, D, μ) satisfies (7) then (f, X, D, μ) satisfies (10).*

Proof. The construction of $Z_{f,X}(y)$ implies that $Z_{f,X}(y) = X^\star$ for all $y \in X^\star$. Therefore, if (f, X, D, μ) satisfies (7) then by taking $y = \bar{x}$ it follows that
$$D_f(\bar{x}, x) \geq D_f(Z_{f,X}(\bar{x}), x) \geq \mu D(Z_{f,X}(\bar{x}), x) = \mu D(\bar{x}, x) \text{ for all } x \in X.$$
□

The following simple example shows that, perhaps contrary to what one might intuitively expect, relative functional growth is not necessarily a relaxation of strong relative convexity unless some additional assumptions are made about f, X, or D.

Example 2. Let $a > 0$ and $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = e^{ax}$. For $X := \mathbb{R}_+$ we have $X^\star = \{0\}$. Thus for $D := D_f$ and $\mu = 1$ the tuple (f, X, D, μ) satisfies (7). However, observe that for all $\hat{\mu} > 0$ and $x \geq 1/(\hat{\mu}a)$
$$f(x) - f^\star = e^{ax} - 1 < \hat{\mu}(1 + axe^{ax}) = \hat{\mu}(f^\star - f'(x)(0 - x)) = \hat{\mu}D(X^\star, x).$$
In particular, $(f, X, D, \hat{\mu})$ does not satisfy (11) for any $\hat{\mu} > 0$.

It can be shown that under additional assumptions on f, X, or D the relative functional growth condition is a relaxation of the relative strong convexity condition. In particular, relative functional growth is a relaxation of relative strong convexity when D is a squared norm as we discuss in Section 4 below.

3 Properties of $L_{f,X,D}$ and $\mu_{f,X,D}$ when f is of the form $g \circ A$

This section develops some properties of the relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ when f is of the form $f := g \circ A$ for $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, and D is bounded in terms of some norm in $\mathbb{R}^n$. The main results of this section are Theorem 1 and Theorem 2. These results provide lower bounds on $\mu_{f,X,D}$ in terms of μ_g and the norms of some canonical set-valued maps that depend on A and X. In a similar vein, Proposition 2 gives an upper bound on $L_{f,X,D}$ in terms of L_g and the norm of a canonical map associated to A and X.

We will rely on the objects $Z_{A,X}(\cdot)$ and $A|C, (A|C)^{-1}$ defined next. For $A \in \mathbb{R}^{m\times n}$, $X \subseteq \mathbb{R}^n$ nonempty and $y \in X$ let

$$Z_{A,X}(y) := \{x \in X : Ax = Ay\}. \tag{13}$$

The set-valued mapping $Z_{A,X} : X \rightrightarrows X$ can be seen as an extension of the set-valued mapping $Z_{f,X} : X \rightrightarrows X$ introduced in Section 2.1.

For $A \in \mathbb{R}^{m\times n}$ and a convex cone $C \subseteq \mathbb{R}^n$ let $A|C : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be the set-valued mapping defined via

$$x \mapsto (A|C)(x) := \begin{cases} \{Ax\} & \text{if } x \in C \\ \emptyset & \text{otherwise.} \end{cases} \tag{14}$$

And let $(A|C)^{-1} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be its inverse, that is,

$$v \mapsto (A|C)^{-1}(v) := \{x \in C : Ax = v\}. \tag{15}$$

Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Define the norms of $A|C$ and of $(A|C)^{-1}$ as follows

$$\|A|C\| := \sup_{\substack{x\in C \\ \|x\|\leq 1}} \|Ax\|, \quad \|(A|C)^{-1}\| := \sup_{\substack{v\in A(C) \\ \|v\|\leq 1}} \inf_{\substack{x\in C \\ Ax=v}} \|x\|. \tag{16}$$

Observe that if $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ is a nonempty convex set such that $A(X)$ contains more than one point then

$$\|A|\operatorname{span}(X - X)\| = \sup_{\substack{y,x\in X \\ x\neq y}} \frac{\|Ay - Ax\|}{\|y - x\|}, \tag{17}$$

where $\operatorname{span}(X - X)$ denotes the linear subspace spanned by $X - X$, that is,

$$\operatorname{span}(X - X) = \{\lambda(x - y) : x, y \in X, \lambda \in \mathbb{R}\}.$$

In particular, the following property of the relative smoothness constant readily follows.

Proposition 2. *Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a nonempty convex set such that $A(X)$ contains more than one point.*

(a) *If $\mathbb{R}^m$ is endowed with the Euclidean norm, $D(y,x) = \frac{1}{2}\|y-x\|^2$ for some norm in $\mathbb{R}^n$, and $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $b \in \mathbb{R}^m$ then*

$$L_{f,X,D} = \|A|\operatorname{span}(X-X)\|^2.$$

(b) *Suppose $\mathbb{R}^m, \mathbb{R}^n$ are endowed with norms and $D(y,x) \geq \frac{1}{2}\|y-x\|^2$ for the norm in $\mathbb{R}^n$. If $f = g \circ A$ where $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is L_g smooth on $A(X)$ for the norm in $\mathbb{R}^m$ then*

$$L_{f,X,D} \leq L_g\|A|\operatorname{span}(X-X)\|^2.$$

Proof. **(a)** This follows from (17) and $D_f(y,x) = \frac{1}{2}\|Ay-Ax\|_2^2$.

(b) This follows from (17) and $D_f(y,x) = D_g(Ay,Ax) \leq \frac{L_g}{2}\|Ay-Ax\|^2$. The latter inequality follows from the L_g smoothness of g.

□

We next discuss far more interesting results that either characterize or lower bound the relative strong convexity constant $\mu_{f,X,D}$.

3.1 Lower bound on $\mu_{f,X,D}$ when X is a convex cone and $A(X)$ is a linear subspace

In this subsection we will consider the special case when $X \subseteq \mathbb{R}^n$ is a convex cone and $A \in \mathbb{R}^{m\times n}$ is such that $A(X)$ is a linear subspace of $\mathbb{R}^m$. The latter condition is equivalent to the following *Slater condition:* there exists $x \in \mathsf{ri}(X)$ such that $Ax = 0$, where $\mathsf{ri}(X)$ denotes the relative interior of X. When this is the case, the norms $\|A|X\|$ and $\|(A|X)^{-1}\|$ have the following geometric interpretation. Let $\mathbb{B}^m$ and $\mathbb{B}^n$ denote the unit balls in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. It is easy to see that if X is a convex cone and $A(X)$ is a linear subspace then

$$\|A|X\| = \inf\{r : A(X \cap \mathbb{B}^n) \subseteq r\mathbb{B}^m \cap A(X)\} \tag{18}$$

and

$$\frac{1}{\|(A|X)^{-1}\|} = \sup\{r : r\mathbb{B}^m \cap A(X) \subseteq A(X \cap \mathbb{B}^n)\}. \tag{19}$$

In other words, $\|A|X\|$ is the radius of the *smallest* ball in $A(X)$ centered at the origin that *contains* $A(X \cap \mathbb{B}^n)$. Similarly, $1/\|(A|X)^{-1}\|$ is the radius of the *largest* ball in $A(X)$ centered at the origin and that is *contained* in $A(X \cap \mathbb{B}^n)$. Example 3 illustrates this geometric interpretation of $\|A|X\|$ and $1/\|(A|X)^{-1}\|$ in a simple instance.

Example 3. Let $A := \begin{bmatrix} 1 & -1 & 0 \\ -\epsilon & -\epsilon & 1 \end{bmatrix}$ for $0 < \epsilon < 1$ and $X = \mathbb{R}^3_+$. Let $\mathbb{R}^2$ be endowed with the Euclidean ℓ_2 norm and let $\mathbb{R}^3$ be endowed with the ℓ_1 norm. In this case $A(X) = \mathbb{R}^2$ and

$$A(X \cap \mathbb{B}^3) = \mathsf{conv}\left\{\begin{bmatrix} 1 \\ -\epsilon \end{bmatrix}, \begin{bmatrix} -1 \\ -\epsilon \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}.$$

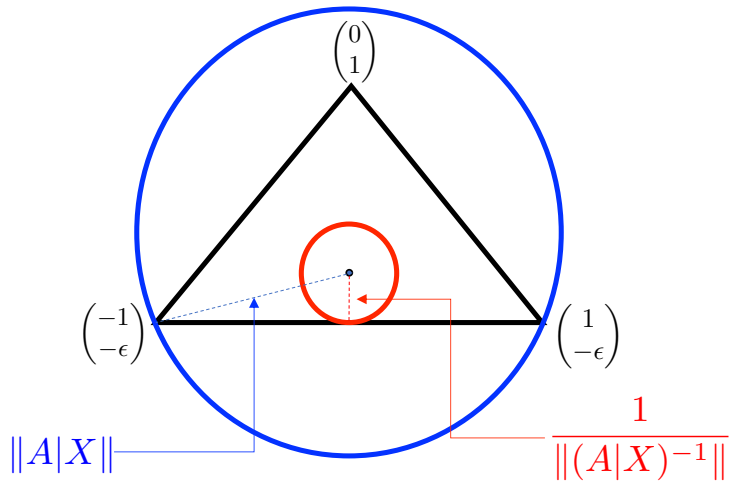

Figure 3: Illustration of $\|A|X\|$ and $1/\|(A|X)^{-1}\|$ for A and X as in Example 3.

Therefore $\|A|X\| = \sqrt{1+\epsilon^2}$ and $1/\|(A|X)^{-1}\| = \epsilon$ as Figure 3 illustrates.

The above norms, especially $\|(A|X)^{-1}\|$ and other related quantities, have been extensively studied in the literature on condition measures for convex optimization [6, 9, 11, 27, 30, 31]. They have been further extended to the broader variational analysis context [7,19]. In particular, when $A(X) = \mathbb{R}^m$ the family of conic systems $Ax = b, x \in X$ is *well-posed*. That is, for all $b \in \mathbb{R}^m$ the conic system $Ax = b, x \in X$ is feasible and remains so for sufficiently small perturbations of (A, b). In this case it follows from [31] that the quantity $1/\|(A|X)^{-1}\|$ is precisely the *distance to ill-posedness* introduced by Renegar [30,31], that is, the size of the smallest perturbation ΔA on A so that the conic system $(A+\Delta A)x = b, x \in X$ is infeasible for some $b \in \mathbb{R}^m$. A similar identity holds for the *distance to non-surjectivity* of closed sublinear set-valued mappings [19]. The latter in turn extends to a far more general identity for the radius of metric regularity [7].

Observe that if $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ is a linear subspace then $A(X)$ is automatically a linear subspace. If in addition $\mathbb{R}^n$ and $\mathbb{R}^m$ are each endowed with Euclidean norms, then (18) and (19) yield

$$\|A|X\| = \sigma_{\max}(A|X) \;\text{ and }\; \frac{1}{\|(A|X)^{-1}\|} = \sigma^+_{\min}(A|X).$$

Corollary 1 and Theorem 1 below show that there is a tight connection between the relative strong convexity constant $\mu_{f,X,D}$ and the norm $\|(A|X)^{-1}\|$ when f is of the form $g \circ A$. Both of these results rely on the following proposition that characterizes a certain type of *Hoffman* constant [15]. Proposition 3 is closely related to developments in [26, 29]. Proposition 3 extends [29, Theorem 2] that only applies to the case $X = \mathbb{R}^n_+$.

Proposition 3. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a convex cone such that $A(X)$ contains more than one point. If $A(X)$ is a linear subspace then*

$$\frac{1}{\|(A|X)^{-1}\|} = \inf_{\substack{x,y\in X\\ x\notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}. \tag{20}$$

Proof. Fix $y \in X$ and $x \in X \setminus Z_{A,X}(y)$. Since $A(X)$ is a linear subspace, it follows that $Ay - Ax \in A(X)$ and thus $Ay - Ax = Au$ for some $u \in X$ with $\|u\| \leq \|(A|X)^{-1}\| \cdot \|Ay-Ax\|$. Hence $x+u \in Z_{A,X}(y)$ and $\|Z_{A,X}(y)-x\| \leq \|u\| \leq \|(A|X)^{-1}\| \cdot \|Ay-Ax\|$. Since this holds for arbitrary $y \in X$ and $x \in X \setminus Z_{A,X}(y)$ we conclude that

$$\frac{1}{\|(A|X)^{-1}\|} \leq \inf_{\substack{y,x\in X\\ x\notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}.$$

To prove the reverse inequality, let $v \in A(X)$ and $0 < \epsilon < \|(A|X)^{-1}\|$ be such that $\|v\| = 1$ and $\|y\| \geq \|(A|X)^{-1}\| - \epsilon$ for all $y \in X$ with $Ay = v$. Pick $\hat{y} \in X$ with $A\hat{y} = v$. Then $\|z\| \geq \|(A|X)^{-1}\| - \epsilon > 0$ for all $z \in Z_{A,X}(\hat{y})$. Thus $\hat{x} := 0 \in X \setminus Z_{A,X}(\hat{y})$ and

$$\frac{1}{\|(A|X)^{-1}\| - \epsilon} \geq \frac{\|A\hat{y} - A\hat{x}\|}{\|Z_{A,X}(\hat{y}) - \hat{x}\|} \geq \inf_{\substack{y,x\in X\\ x\notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}.$$

To finish let $\epsilon \to 0$. □

Proposition 3 readily yields the following result that generalizes Example 1.

Corollary 1. *Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm $\|\cdot\|_2$, $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$, and $D(x,y) = \frac{1}{2}\|x-y\|^2$. If $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, $X \subseteq \mathbb{R}^n$ is a convex cone, and $A(X)$ is a linear subspace that contains more than one point then*

$$\mu_{f,X,D} = \frac{1}{\|(A|X)^{-1}\|^2}.$$

Proof. This follows from Proposition 3 and the observation that for this choice of f and X we have $Z_{f,X}(y) = Z_{A,X}(y)$ and $f(y) - f(x) - \langle \nabla f(x), y-x\rangle = \frac{1}{2}\|Ay-Ax\|_2^2$. □

The following result extends Corollary 1 to a broader class of functions.

Theorem 1. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x,y) \leq \frac{1}{2}\|x-y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function, and $X \subseteq \mathbb{R}^n$ be a convex cone such that $A(X)$ is a linear subspace that contains more than one point. If g is μ_g strongly convex on $A(X)$ for the norm $\|\cdot\|$ in $\mathbb{R}^m$ then the function $f = g \circ A$ satisfies*

$$\mu_{f,X,D} \geq \frac{\mu_g}{\|(A|X)^{-1}\|^2}.$$

Proof. Observe that $D_f(y,x) = g(Ay) - g(Ax) - \langle g(Ax), A(y-x)\rangle$ for all $y, x \in X$ Since g is μ_g strongly convex, it follows that $D_f(y,x) \geq \mu_g \|Ay - Ax\|^2/2$ for all $y, x \in X$ and $Z_{f,X}(y) = \{x \in X : Ax = Ay\} = Z_{A,X}(y)$ for all $y \in X$. Therefore Proposition 3 implies that

$$\mu_{f,X,D} \geq \inf_{\substack{y,x\in X \\ x\notin Z_{A,X}(y)}} \frac{D_f(y,x)}{\|Z_{A,X}(y) - x\|^2/2} \geq \inf_{\substack{y,x\in X \\ x\notin Z_{A,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2} = \frac{\mu_g}{\|(A|X)^{-1}\|^2}.$$

□

If f, X, D are as in Corollary 1 then by Proposition 2 the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ is

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} = \left(\|A|\operatorname{span}(X)\| \cdot \|(A|X)^{-1}\|\right)^2$$

which has a striking resemblance to the classical condition number (1) of $f(x) = \frac{1}{2}\|Ax - b\|_2^2$. More generally, if f, X, D are as in Theorem 1 and g is also L_g smooth then by Proposition 2 we obtain the following bound on the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ in terms of the condition number of g and a condition number of the pair (A, X):

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} \leq \frac{L_g}{\mu_g} \cdot \left(\|A|\operatorname{span}(X)\| \cdot \|(A|X)^{-1}\|\right)^2. \tag{21}$$

3.2 Lower bound on $\mu_{f,X,D}$ when X is a polyhedron

The results in Section 3.1 require X to be a convex cone and $A(X)$ to be a linear subspace. We next provide some results of similar flavor that relax these assumptions in exchange for the assumption that X is a polyhedron. The crux of the main results in this section is Proposition 4. This technical result is drawn from the recent paper of Peña, Vera, and Zuluaga [26]. The latter paper develops a number of properties of a new class of *relative Hoffman bounds*. In particular, it introduces the sets of tangent cones $\mathcal{T}(X)$ and $\mathcal{T}(A|X)$ described below. These two sets of tangent cones are at the heart of the main developments in [26].

For a nonempty polyhedron $X \subseteq \mathbb{R}^n$ let $\mathcal{T}(X) := \{T_X(x) : x \in X\}$, where $T_X(x)$ is the tangent cone of X at x, that is,

$$T_X(x) := \{d \in \mathbb{R}^n : x + td \in X \ \text{ for some } \ t > 0\}.$$

We will rely on the following subset of $\mathcal{T}(X)$ that depends on the how A and X fit together. Let

$$\mathcal{T}(A|X) := \{C \in \mathcal{T}(X) : A(C) \text{ is a linear subspace and } C \text{ is minimal}\}. \tag{22}$$

In this definition, *minimal* is to be interpreted as minimal with respect to inclusion. This restriction guarantees that the set $\mathcal{T}(A|X)$ is of minimal size as it does not include redundant cones from $\mathcal{T}(X)$.

Observe that $\mathcal{T}(X)$ is finite since X is polyhedral and thus $\mathcal{T}(A|X)$ is finite as well. The following example illustrates the interesting relationship between A and the tangent cones of X captured by $\mathcal{T}(A|X)$.

Example 4. Suppose $A \in \mathbb{R}^{m\times n}$ and $X = \mathbb{R}^n_+$. In this case each element of $\mathcal{T}(X)$ is of the form $C_I = \{x \in \mathbb{R}^n : x_I \geq 0\}$ for some $I \subseteq \{1, \ldots, n\}$. Observe that $A(C_I)$ is a linear subspace if and only if $Ax = 0$, $x_I > 0$ is feasible. Thus the set $\mathcal{T}(A|X)$ is in one-to-one correspondence with the maximal sets $I \subseteq \{1, \ldots, n\}$ such that $Ax = 0$, $x_I > 0$ is feasible.

Observe that $\mathcal{T}(A|X) = \{X\}$ when X is a polyhedral cone and $A(X)$ is a linear subspace. Thus the following proposition subsumes Proposition 3 when X is polyhedral.

Proposition 4. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Then*

$$\min_{C\in\mathcal{T}(A|X)} \frac{1}{\|(A|C)^{-1}\|} = \inf_{\substack{y,x\in X \\ x\notin Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|}. \tag{23}$$

Proof. This follows as a special case of [26, Proposition 5 and Corollary 3]. □

Corollary 2. *Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm $\|\cdot\|_2$, $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$, and $D(x, y) = \frac{1}{2}\|x - y\|^2$. If $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and $X \subseteq \mathbb{R}^n$ is a polyhedron such that $A(X)$ contains more than one point then*

$$\mu_{f,X,D} = \min_{C\in\mathcal{T}(A|X)} \frac{1}{\|(A|C)^{-1}\|^2}.$$

Proof. Proceed exactly as in the proof of Corollary 1 but apply Proposition 4 instead of Proposition 3. □

Theorem 2. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x, y) \leq \frac{1}{2}\|x - y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function, and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. If g is μ_g strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ then the function $f = g \circ A$ satisfies*

$$\mu_{f,X,D} \geq \min_{C\in\mathcal{T}(A|X)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Proof. Proceeding exactly as in the proof of Theorem 1 but applying Proposition 4 instead of Proposition 3 we get

$$\mu_{f,X,D} \geq \inf_{\substack{y,x\in X \\ x\notin Z_{A,X}(y)}} \frac{D_f(y, x)}{\|Z_{A,X}(y) - x\|^2/2} \geq \inf_{\substack{y,x\in X \\ x\notin Z_{A,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2} = \min_{C\in\mathcal{T}(A|X)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

□

Observe that if X is polyhedral then $\operatorname{span}(X - X) \in \mathcal{T}(X)$ and

$$\|A|\operatorname{span}(X - X)\| = \max_{C \in \mathcal{T}(X)} \|A|C\|.$$

Thus Proposition 2 implies that for f, X, D as in Corollary 2, the relative condition $L_{f,X,D}/\mu_{f,X,D}$ has the following expression, which is again strikingly similar to the classical condition number (1) of $f(x) = \frac{1}{2}\|Ax - b\|_2^2$:

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} = \left(\max_{C \in \mathcal{T}(X)} \|A|C\| \cdot \max_{C \in \mathcal{T}(A|X)} \|(A|C)^{-1}\| \right)^2.$$

Proposition 2 also implies that if f, X, D are as in Theorem 2 and g is L_g smooth then the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be bounded in terms of the condition number of g and a condition number of the pair (A, X) as follows:

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} \leq \frac{L_g}{\mu_g} \cdot \left(\max_{C \in \mathcal{T}(X)} \|A|C\| \cdot \max_{C \in \mathcal{T}(A|X)} \|(A|C)^{-1}\| \right)^2. \tag{24}$$

We next place some of the developments by Peña and Rodríguez [28] in the context of this paper. To that end, consider the special case when X is the standard simplex $\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$ in $\mathbb{R}^n$. For $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ let $\mathsf{conv}(A) := \mathsf{conv}(\{a_1, \ldots, a_n\}) = \{Ax : x \in \Delta_{n-1}\}$ and let $\mathsf{faces}(\mathsf{conv}(A))$ denote the set of faces of $\mathsf{conv}(A)$. Furthermore, for $F \in \mathsf{faces}(\mathsf{conv}(A))$ let $A \setminus F$ denote the set of columns of A that do not belong to F. Suppose $\mathbb{R}^m$ is endowed with a norm and for $F, G \subseteq \mathbb{R}^m$ let $\mathsf{dist}(F, G) := \inf_{u \in F, v \in G} \|u - v\|$. Following [28] define the *facial distance* $\Phi(A)$ of A as follows

$$\Phi(A) := \min_{\substack{F \in \mathsf{faces}(\mathsf{conv}(A)) \\ \emptyset \neq F \neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)). \tag{25}$$

Let $\mathsf{diam}(A)$ denote the *diameter* of the set of columns of A defined as follows

$$\mathsf{diam}(A) := \max_{i,j \in \{1,\ldots,n\}} \|a_i - a_j\|. \tag{26}$$

In the special case when $X = \Delta_{n-1}$ it follows from [28, Theorem 1] that (23) in Proposition 4 has the following geometric characterization

$$\min_{\substack{y, x \in \Delta_{n-1} \\ x \notin Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|_1} = \frac{\Phi(A)}{2}. \tag{27}$$

Furthermore, in this same special case when $X = \Delta_{n-1}$ it is easy to see that (17) has the following geometric characterization

$$\max_{\substack{x, y \in \Delta_{n-1} \\ x \neq y}} \frac{\|Ay - Ax\|}{\|y - x\|_1} = \frac{\mathsf{diam}(A)}{2}. \tag{28}$$

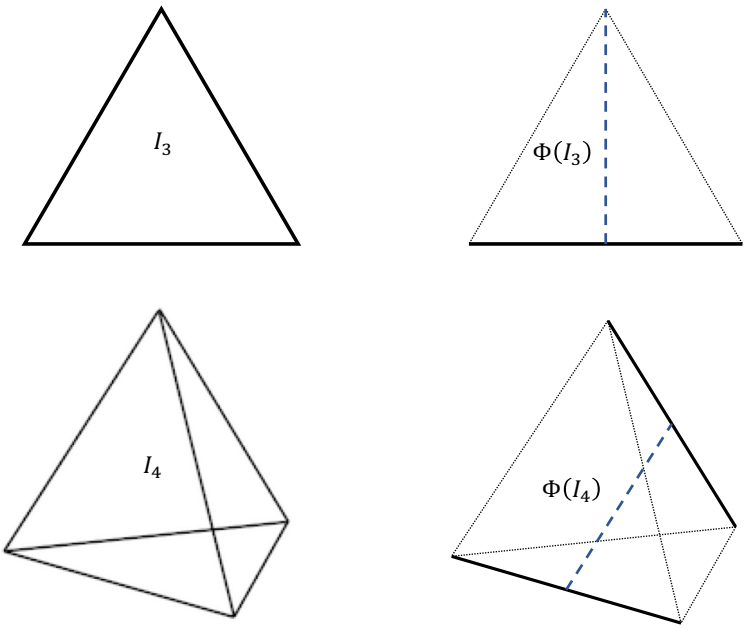

Figure 4: Depiction of $\mathsf{conv}(A)$ and $\Phi(A)$ for $A = I_3$ and $A = I_4$.

Figure 4 gives a visualization of the facial distance $\Phi(A)$ for $A = I_3$ and $A = I_4$. It depicts $\mathsf{conv}(A)$ and $\Phi(A)$ in the hyperplane $\{x : \langle \mathbf{1}, x\rangle = 1\}$.

Example 5 below, a special case of Corollary 2, shows that for $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $X = \Delta_{n-1}$, and $D(y, x) = \frac{1}{2}\|y - x\|_1^2$ the relative condition number $L_{f,\Delta_{n-1},D}/\mu_{f,\Delta_{n-1},D}$ is the square of $\mathsf{diam}(A)/\Phi(A)$, which has a flavor of an aspect ratio of $\mathsf{conv}(A)$. This gives an interesting analogy to (1).

Example 5. Suppose $\mathbb{R}^n$ is endowed with the one-norm, $\mathbb{R}^m$ is endowed with the Euclidean norm, and $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ with at least two different columns and $b \in \mathbb{R}^m$. Then for $D(y, x) := \frac{1}{2}\|y - x\|_1^2$ Corollary 2 and identities (28) and (27) yield

$$L_{f,\Delta_{n-1},D} = \frac{\mathsf{diam}(A)^2}{4} \quad \text{and} \quad \mu_{f,\Delta_{n-1},D} = \frac{\Phi(A)^2}{4}.$$

In particular,

$$\frac{L_{f,\Delta_{n-1},D}}{\mu_{f,\Delta_{n-1},D}} = \left(\frac{\mathsf{diam}(A)}{\Phi(A)}\right)^2.$$

More generally, if $f(x) = g(Ax)$ for some L_g smooth and μ_g strongly convex function g then

$$L_{f,\Delta_{n-1},D} \le \frac{L_g \cdot \mathsf{diam}(A)^2}{4} \quad \text{and} \quad \mu_{f,\Delta_{n-1},D} \ge \frac{\mu_g \cdot \Phi(A)^2}{4}.$$

In particular,

$$\frac{L_{f,\Delta_{n-1},D}}{\mu_{f,\Delta_{n-1},D}} \leq \frac{L_g}{\mu_g} \cdot \left(\frac{\mathsf{diam}(A)}{\Phi(A)}\right)^2.$$

4 Properties of $\mu^\star_{f,X,D}$, and $\mu^\sharp_{f,X,D}$ when D is a squared norm

We next provide bounds on $\mu^\star_{f,X,D}$ and $\mu^\sharp_{f,X,D}$ analogous to those developed in Section 3 for $\mu_{f,X,D}$. Proposition 1 already established $\mu^\star_{f,X,D} \geq \mu_{f,X,D} \geq 0$. It is intuitively clear that $\mu^\star_{f,X,D}$ could be a lot larger. When D is a squared norm, the exact same technique used in [22, Theorem 1] show that $\mu^\sharp_{f,X,D} \geq \mu^\star_{f,X,D}$. Indeed, when D is a squared norm, the relationship among other variants of strong convexity introduced [22] extend to our context in a straightforward fashion as we next explain.

Definition 4. Suppose (f, X, D) satisfy Assumption 2.

(a) We say that f has D-*under approximation* on X if there exists a constant $\mu > 0$ such that

$$D_f(x, \bar{x}) \geq \mu D(\bar{x}, x) \;\text{ for all }\; x \in X. \tag{29}$$

(b) We say that f has D-*gradient growth* on X if there exists a constant $\mu > 0$ such that

$$\langle \nabla f(x) - \nabla f(\bar{x}), x - \bar{x} \rangle \geq \mu D(\bar{x}, x) \;\text{ for all }\; x \in X. \tag{30}$$

Suppose (f, X, D) satisfies Assumption 2 and D is a squared norm. Then for $\mu > 0$ [22, Theorem 4] yields the following chain of implications for (f, X, D, μ):

$$(7) \Rightarrow (10) \Rightarrow (29) \Rightarrow (30) \Rightarrow (11).$$

We note that [22, Theorem 4] is stated and proven for the Euclidean norm but the same statement and proof hold for any norm.

From the above chain of implications it follows that if (f, X, D) satisfies Assumption 2 and D is a squared norm then $\mu_{f,X,D} \leq \mu^\star_{f,X,D} \leq \mu^\sharp_{f,X,D}$. In particular, any lower bound on $\mu_{f,X,D}$, such as those in Theorem 1 or Theorem 2, is also a lower bound on $\mu^\star_{f,X,D}$ and on $\mu^\sharp_{f,X,D}$ when D is a squared norm. We next show that the ideas in Section 3 can be extended to obtain sharper bounds on these two constants.

4.1 A sharper lower bound on $\mu^\star_{f,X,D}$

Suppose $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ is a polyhedron such that $A(X)$ contains more than one point, and $S \subseteq X$ is nonempty. Proposition 4 readily implies

$$\inf_{\substack{y\in S,\, x\in X \\ x\notin Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|} \geq \min_{C\in\mathcal{T}(A|X)} \frac{1}{\|(A|C)^{-1}\|} > 0. \tag{31}$$

Proposition 5 below, which extends Proposition 4, gives a sharper version of (31). Suppose $A \in \mathbb{R}^{m\times n}$, $X \subseteq \mathbb{R}^n$ is a polyhedron, and $S \subseteq X$ is nonempty. Let

$$\mathcal{T}(A|X,S) := \{T_X(x;A,S) : x \in X\} \tag{32}$$

where

$$T_X(x;A,S) := \{d \in \mathbb{R}^n : x + td \in X \text{ and } A(x+td) \in \mathsf{conv}(A(S)) \text{ for some } t > 0\}.$$

Proposition 5 can be proven via a straightforward modification of techniques in [26]. We provide the details of this modification in Appendix A.

Proposition 5. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Then for all nonempty $S \subseteq X$*

$$\inf_{\substack{y\in S,\, x\in X\\ x\notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|} \geq \inf_{C\in\mathcal{T}(A|X,S)} \frac{1}{\|(A|C)^{-1}\|} \geq \min_{C\in\mathcal{T}(A|X)} \frac{1}{\|(A|C)^{-1}\|}. \tag{33}$$

Furthermore, if $A(S)$ is convex then

$$\inf_{\substack{y\in S,\, x\in X\\ x\notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|} = \inf_{C\in\mathcal{T}(A|X,S)} \frac{1}{\|(A|C)^{-1}\|}. \tag{34}$$

Corollary 3. *Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm $\|\cdot\|_2$, $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$, and $D(x,y) = \frac{1}{2}\|x-y\|^2$. If $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and $X \subseteq \mathbb{R}^n$ is a polyhedron such that $A(X)$ contains more than one point and $X^\star := \operatorname{argmin}_{x\in X} f(x) \neq \emptyset$. Then*

$$\mu^\star_{f,X,D} = \inf_{C\in\mathcal{T}(A|X,X^\star)} \frac{1}{\|(A|C)^{-1}\|^2}.$$

Proof. Proceed exactly as in the proof of Corollary 1 but apply Proposition 5 instead of Proposition 3. □

The following theorem gives a lower bound on $\mu^\star_{f,X,D}$ analogous to the one on $\mu_{f,X,D}$ in Theorem 2. In light of Proposition 5, the lower bound on $\mu^\star_{f,X,D}$ in Theorem 3 is at least as large, and possibly much larger, than the one on $\mu_{f,X,D}$ in Theorem 2.

Theorem 3. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(y,x) \leq \frac{1}{2}\|y-x\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ has more than one point. If g is μ_g-strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ then the function $f = g\circ A$ satisfies*

$$\mu^\star_{f,X,D} \geq \inf_{C\in\mathcal{T}(A|X,X^\star)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Proof. Observe that for all $y \in X^\star$ and $x \in X$

$$D_f(y,x) = g(Ay) - g(Ax) - \langle g(Ax), A(y-x)\rangle .$$

Since g is μ_g strongly convex on $A(X)$, it follows that $D_f(y,x) \geq \mu_g \|Ay - Ax\|^2/2$ for all $y \in X^\star$ and $x \in X$, and it also follows that $Z_{A,X}(y) = \{x \in X : Ax = Ay\} = X^\star$ for all $y \in X^\star$. Therefore

$$\mu^\star_{f,X,D} \geq \inf_{x \in X \setminus X^\star} \frac{D_f(\bar{x},x)}{\|\bar{x}-x\|^2/2} \geq \inf_{\substack{y \in X^\star \\ x \in X \setminus X^\star}} \frac{D_f(y,x)}{\|y-x\|^2/2} \geq \inf_{\substack{y \in X^\star,\, x \in X \\ x \notin Z_{A,X}(y)}} \frac{\mu_g \|Ay-Ax\|^2}{\|Z_{A,X}(y)-x\|^2}.$$

To finish, apply Proposition 5. □

Once again there is an interesting connection with the developments in [28] when $X = \Delta_{n-1}$. Consider the special case when $X = \Delta_{n-1}$, $A \in \mathbb{R}^{m \times n}$ has at least two different columns, $S \subseteq \Delta_{n-1}$ is nonempty, and $G \in \mathsf{faces}(\mathsf{conv}(A))$ is the smallest face of $\mathsf{conv}(A)$ that contains $A(S)$. From [28, Theorem 3] it follows that if $\mathbb{R}^n$ is endowed with the one-norm then

$$\inf_{\substack{y \in S, x \in X \\ x \notin Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|_1} \geq \min_{\substack{F \in \mathsf{faces}(G) \\ \emptyset \neq F \neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)). \tag{35}$$

The following example illustrates the difference between $\mu_{f,X,D}$ and $\mu^\star_{f,X,D}$.

Example 6. Suppose $\mathbb{R}^n$ is endowed with the one-norm and $D(y,x) := \frac{1}{2}\|y-x\|_1^2$. Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm, and $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m \times n}$ with at least two different columns and $b \in \mathbb{R}^m$. As noted in Example 5, in this case

$$\mu_{f,\Delta_{n-1},D} = \frac{\Phi(A)^2}{4} = \frac{1}{4}\left(\min_{\substack{F \in \mathsf{faces}(\mathsf{conv}(A)) \\ \emptyset \neq F \neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)) \right)^2 .$$

This relative strong convexity constant depends only on A but not on b. On the other hand, the smallest face of $\mathsf{conv}(A)$ containing $X^\star$ is

$$G(b) := \operatorname*{argmin}_{G \in \mathsf{faces}(\mathsf{conv}(A))} \mathsf{dist}(G, b),$$

which evidently depends on both A and b. Theorem 3 and (35) yield

$$\mu^\star_{f,\Delta_{n-1},D} \geq \frac{1}{4}\left(\min_{\substack{F \in \mathsf{faces}(G(b)) \\ \emptyset \neq F \neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)) \right)^2 .$$

It is evident that

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

Furthermore, as it is illustrated in [28], the difference between these two quantities can be arbitrarily large. Consequently, the bound in Theorem 3 can be far sharper than that in Theorem 2.

4.2 A sharper lower bound on $\mu^\sharp_{f,X,D}$

Suppose $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is defined as $f(x) = g(Ax)+\langle c, x\rangle$ where $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ is a strongly convex function, $A \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^n$. Theorem 3 does not apply to this kind of function due to the extra linear term $\langle c, x\rangle$. Indeed for a function of this form the constant $\mu^\star_{f,X,D}$ may be zero, see Example 7 below. On the other hand, the next result shows that for a function of this form and for a polyhedral set X it is always the case that $\mu^\sharp_{f,X,D} > 0$ provided a suitable linear cut is added to X.

Theorem 4. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x, y) \le \frac{1}{2}\|x - y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Suppose $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ is μ_g-strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ and $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is defined via $f(x) = g(Ax)+\langle c, x\rangle$. Then the vector $v := 2\nabla f(y)$ is the same for all $y \in X^\star$ and satisfies $\langle v, x - y\rangle \ge 0$ for all $x \in X$, $y \in X^\star$. Furthermore, one of the following two possible cases applies depending on the range of values of $\langle v, x - y\rangle$ for $x \in X, y \in X^\star$.*

Case 1: *For all $x \in X, y \in X^\star$ we have $\langle v, x - y\rangle = 0$. In this case*

$$\mu^\sharp_{f,X,D} \ge \inf_{C\in\mathcal{T}(A|X,X^\star)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Case 2: *For some $x \in X, y \in X^\star$ we have $\langle v, x - y\rangle > 0$. In this case for all $\delta > 0$*

$$\mu^\sharp_{f,X_\delta,D} \ge \inf_{C\in\mathcal{T}(M|X_\delta,X^\star)} \frac{1}{\|(M|C)^{-1}\|^2},$$

for the polyhedron $X_\delta := \{x \in X : \langle v, x - y\rangle \le \delta \text{ for all } y \in X^\star\} \supseteq X^\star$, the matrix $M \in \mathbb{R}^{(m+1)\times n}$, and the norm $\|\cdot\|$ in $\mathbb{R}^{m+1}$ defined as follows

$$M := \begin{bmatrix} \sqrt{\mu_g}\cdot A \\ \frac{1}{\sqrt{\delta}}\cdot v^\mathsf{T} \end{bmatrix} \quad \text{and} \quad \left\|\begin{bmatrix} y \\ y_{m+1}\end{bmatrix}\right\| := \sqrt{\|y\|^2 + y_{m+1}^2}.$$

Proof. The optimality conditions for $\min_{x\in X} f(x)$ imply that

$$\langle \nabla f(y), x - y\rangle = \left\langle A^\mathsf{T}\nabla g(Ay) + c, x - y\right\rangle \ge 0 \text{ for all } x \in X, y \in Y^\star. \tag{36}$$

Thus for all $y, y' \in X^\star$ the strong convexity of g and (36) imply

$$\mu_g\|Ay - Ay'\|^2 \le \langle \nabla g(Ay) - \nabla g(Ay'), Ay - Ay'\rangle = \langle \nabla f(y) - \nabla f(y'), y - y'\rangle \le 0.$$

Hence $Ay = Ay'$ whenever $y, y' \in X^\star$. In particular, $v = 2\nabla f(y) = 2(A^\mathsf{T}\nabla g(Ay) + c)$ is the same for all $y \in X^\star$. Furthermore, the optimality conditions for $\min_{x\in X} f(x)$ imply that $\langle v, x - y\rangle \ge 0$ for all $x \in X, y \in Y^\star$. In particular, $\langle v, y\rangle = \min_{x\in X}\langle v, x\rangle$ for all $y \in X^\star$.

Next, the strong convexity of g on $A(X)$ implies that for all $x \in X, y \in X^\star$

$$\begin{aligned} f(x) - f^\star &= g(Ax) - g(Ay) + \langle c, x - y\rangle \\ &\geq \frac{\mu_g}{2}\|Ax - Ay\|^2 + \langle \nabla g(Ay), Ax - Ay\rangle + \langle c, x - y\rangle \\ &= \frac{1}{2}\left(\mu_g\|Ax - Ay\|^2 + \langle v, x - y\rangle\right). \end{aligned}$$

If $\langle v, x - y\rangle = 0$ for all $x \in X, y \in X^\star$ then Case 1 applies. In this case $Z_{A,X}(y) = \{x \in X : Ax = Ay\} = X^\star$ for all $y \in X^\star$ and thus

$$\mu^\sharp_{f,X,D} = \inf_{\substack{y\in X^\star \\ x\in X\setminus X^\star}} \frac{f(x) - f^\star}{\|y - x\|^2/2} \geq \inf_{\substack{y\in X^\star,\, x\in X \\ x\notin Z_{A,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2}.$$

If $\langle v, x - y\rangle > 0$ for some $x \in X, y \in X^\star$ then Case 2 applies. In this case $Z_{M,X}(y) = \{x \in X : Ax = Ay, \langle v, x\rangle = \langle v, y\rangle\} = X^\star$ for all $y \in X^\star$ and thus

$$\mu^\sharp_{f,X_\delta,D} = \inf_{\substack{y\in X^\star \\ x\in X_\delta\setminus X^\star}} \frac{f(x) - f^\star}{\|y - x\|^2/2} \geq \inf_{\substack{y\in X^\star,\, x\in X_\delta \\ x\notin Z_{M,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2 + \langle v, y - x\rangle}{\|Z_{M,X}(y) - x\|^2}.$$

Next, observe that for $y \in X^\star$ and $x \in X_\delta$

$$\mu_g\|Ay - Ax\|^2 + \langle v, y - x\rangle \geq \mu_g\|Ay - Ax\|^2 + \frac{\langle v, y - x\rangle^2}{\delta} = \|My - Mx\|^2.$$

To finish, apply Proposition 5 in either case. □

Observe that if X in Theorem 4 is bounded then Case 2 gives a lower bound on $\mu^\sharp_{f,X,D}$ by taking $\delta := \max_{x\in X, y\in X^\star}\langle v, x - y\rangle$ because $X = X_\delta$ for this choice of δ.

We conclude this section with a simple example showing that $\mu^\sharp_{f,X,D} > \mu^\star_{f,X,D} = 0$ can occur. The example also shows that the additional bound on X_δ in Theorem 4, Case 2 cannot simply dropped without making some additional assumptions.

Example 7. Let $\mathbb{R}^3$ be endowed with the one-norm and let $D(y, x) := \frac{1}{2}\|y - x\|_1^2$. Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is as follows

$$f(x) = \frac{1}{2}(x_1 - x_2)^2 + x_3.$$

If $X := \Delta_2 \subseteq \mathbb{R}^3$ then $X^\star = \{\begin{bmatrix}1/2 & 1/2 & 0\end{bmatrix}^\mathsf{T}\}$. For $x = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^\mathsf{T}$ we have $f(\bar{x}) - f(x) - \langle \nabla f(x), \bar{x} - x\rangle = 0$ and $\|\bar{x} - x\|_1 = 2$. Hence $\mu^\star_{f,X,D} = 0$. On the other hand, Theorem 4 implies that $\mu^\sharp_{f,X,D} > 0$. A more careful calculation shows that in this case $\mu^\sharp_{f,X,D} = 1/2$.

On the other hand, if $X = \mathbb{R}^3_+$ then $X^\star = \{\begin{bmatrix}t & t & 0\end{bmatrix}^\mathsf{T} : t \geq 0\}$. For $t > 0$ and $x = \begin{bmatrix}0 & 0 & t\end{bmatrix}^\mathsf{T}$ we have $f(x) - f^\star = t$ and $\|X^\star - x\|_1 = t$. Therefore $\mu^\sharp_{f,X,D} = 0$. Furthermore, in the context of Theorem 4 we have $v = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^\mathsf{T}$. Thus for all $\delta > 0$ we have $X_\delta := \{x \in X : x_3 \leq \delta\}$ and $\mu^\sharp_{f,X_\delta,D} = 2/\delta > 0$.

5 Convergence of first-order methods

This section details linear convergence results for the mirror descent algorithm, Frank-Wolfe algorithm, and Frank-Wolfe algorithm with away steps for problem (2). The linear convergence statements for the three algorithms are strikingly similar. They are stated in terms of the relative constants $L_{f,X,D}$ and $\mu^{\star}_{f,X,D}, \mu^{\sharp}_{f,X,D}$ for suitable choices of distance-like functions D.

5.1 Mirror descent algorithm

Suppose $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable on $X \subseteq \mathbb{R}^n$ and the *Bregman proximal map*

$$g \mapsto \operatorname*{argmin}_{y \in X} \{\langle g, y\rangle + L D_h(y, x)\}$$

is computable for $x \in X$ and $L > 0$. The *mirror descent algorithm* for problem (2) is based on the following update for $x \in X$:

$$x_+ := \operatorname*{argmin}_{y \in X} \{\langle \nabla f(x), y\rangle + L D_h(y, x)\}.$$

Algorithm 1 gives a description of the mirror descent algorithm for (2).

Algorithm 1 Mirror descent algorithm

1: Pick $x_0 \in X$;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: choose $L_k > 0$
4: $x_{k+1} = \operatorname*{argmin}_{y \in X} \{\langle \nabla f(x_k), y\rangle + L_k D_h(y, x_k)\}$
5: **end for**

Proposition 6 and Proposition 7 show the linear convergence of Algorithm 1 provide suitable relative smoothness and relative quasi-strong convexity or relative functional growth conditions hold. Throughout the remaining of this section we assume that (f, X, D_h) satisfy Assumption 1.

We should note that Proposition 6 and its proof are straightforward modifications of the linear convergence results in [20, 32]. However, Proposition 6 shows that the linear convergence of Algorithm 1 holds with a sharper rate and under more general assumptions than those in [20, 32]. In particular, the rates in Proposition 6 is stated in terms of a relative quasi-strong convexity constant, which is always at least as large and possibly much larger than the kind of relative strong convexity constant in [20,32]. Furthermore, our results in Section 3 and Section 4 guarantee linear convergence when f is of the form $g \circ A$ provided g and h satisfy smoothness and strong convexity assumptions. The linear convergence results in [20, 32] do not apply for functions of this form because they are not strictly convex and thus the kind of relative strong convexity constant in [20, 32] is typically zero.

The following lemma, which is a straightforward extension of results presented in [32], provides the crux of the proof of Proposition 6.

Lemma 1. *Suppose* $L := L_{f,X,D_h} < \infty$ *and* $\mu := \mu^\star_{f,X,D_h} > 0$. *If* $x \in X$ *and*

$$x_+ = \operatorname*{argmin}_{y \in X} \{f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x)\} \tag{37}$$

then

$$f(x_+) - f^\star \leq (L - \mu) D_h(\bar{x}, x) - L D_h(\bar{x}, x_+). \tag{38}$$

Proof. Since $L = L_{f,X,D_h}$ and $\mu = \mu^\star_{f,X,D_h}$ we have

$$f(x_+) \leq f(x) + \langle \nabla f(x), x_+ - x \rangle + L D_h(x_+, x). \tag{39}$$

and

$$f(x) \leq f^\star + \langle \nabla f(x), x - \bar{x} \rangle - \mu D_h(\bar{x}, x). \tag{40}$$

In addition, the three-point property of D_h [5, Lemma 3.1] yields

$$D_h(x_+, x) = D_h(\bar{x}, x) - D_h(\bar{x}, x_+) + \langle \nabla h(x_+) - \nabla h(x), x_+ - \bar{x} \rangle . \tag{41}$$

By putting together (39), (40), and (41) we get

$$\begin{aligned} f(x_+) \leq f^\star &+ (L - \mu) D_h(\bar{x}, x) - L D_h(\bar{x}, x_+) \\ &+ \langle \nabla f(x) + L(\nabla h(x_+) - \nabla h(x)), x_+ - \bar{x} \rangle . \end{aligned}$$

We get (38) by observing that the optimality conditions for (37) imply

$$\langle \nabla f(x) + L(\nabla h(x_+) - \nabla h(x)), x_+ - \bar{x} \rangle \leq 0.$$

□

Proposition 6. *Suppose* $L := L_{f,X,D_h} < \infty$ *and* $\mu := \mu^\star_{f,X,D_h} > 0$. *If* $L_k = L$, $k = 0, 1, \ldots$ *in Algorithm 1 then the iterates generated by Algorithm 1 satisfy*

$$D_h(X^\star, x_k) \leq \left(1 - \frac{\mu}{L}\right)^k D_h(X^\star, x_0) \quad \textit{for} \quad k = 0, 1, \ldots \tag{42}$$

and

$$f(x_k) - f^\star \leq L \left(1 - \frac{\mu}{L}\right)^k D_h(X^\star, x_0) \quad \textit{for} \quad k = 1, 2, \ldots .$$

Proof. Lemma 1 applied to $x = x_k$ implies that

$$(L - \mu) D_h(\bar{x}_k, x_k) - L D_h(\bar{x}_k, x_{k+1}) \geq f(x_{k+1}) - f^\star \geq 0 \quad \text{for} \quad k = 0, 1, \ldots . \tag{43}$$

Therefore

$$D_h(X^\star, x_{k+1}) \leq D_h(\bar{x}_k, x_{k+1}) \leq \left(1 - \frac{\mu}{L}\right) D_h(\bar{x}_k, x_k) \quad \text{for} \quad k = 0, 1, \ldots .$$

Thus (42) readily follows. Inequality (43) also yields

$$f(x_k) - f^\star \leq L \left(1 - \frac{\mu}{L}\right) D_h(X^\star, x_{k-1}) \leq L \left(1 - \frac{\mu}{L}\right)^k D_h(X^\star, x_0) \quad \text{for} \quad k = 1, 2, \ldots .$$

□

Proposition 6 implies that if $f \in q\mathcal{S}_{L,\mu}(X, D_h)$ then Algorithm 1 yields $x_k \in X$ such that $f(x_k) - f^\star < \epsilon$ in at most

$$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{LD_h(X^\star, x_0)}{\epsilon}\right)\right)$$

iterations.

Proposition 7 below shows that the same kind of iteration bound holds under a relative functional growth assumption instead of the quasi strong convexity assumption in Proposition 6. We note that although Proposition 7 is similar in flavor to Proposition 6, it is stated in terms of the novel concept of relative functional growth. Furthermore, neither Proposition 6 nor Proposition 7 implies the other since neither $\mu^\star_{f,X,D_h}$ nor $\mu^\sharp_{f,X,D_h}$ necessarily bounds the other. (See Example 2 and Example 7.)

Proposition 7. *Suppose $L := L_{f,X,D_h} < \infty$ and $\mu := \mu^\sharp_{f,X,D_h} > 0$. If $L_k = L,\ k = 0, 1, \ldots$ in Algorithm 1 then for $K = \lceil 2L/\mu \rceil$ the iterates generated by Algorithm 1 satisfy*

$$D_h(X^\star, x_{k+K}) \leq \frac{D_h(X^\star, x_k)}{2} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{44}$$

In addition, Algorithm 1 yields $x_k \in X$ such that $f(x_k) - f^\star < \epsilon$ in at most

$$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{LD_h(X^\star, x_0)}{\epsilon}\right)\right) \tag{45}$$

iterations.

Proof. Since $f \in \mathcal{F}_{L,\mu}(X, D_h)$ and $L_k = L$, it follows from [20, Theorem 3.1] that the $(k+K)$-th iterate generated by Algorithm 1 satisfies

$$D_h(X^\star, x_{k+K}) \leq \frac{1}{\mu}(f(x_{k+K}) - f^\star) \leq \frac{L}{\mu K} D_h(X^\star, x_k) \leq \frac{D_h(X^\star, x_k)}{2}.$$

Thus (44) follows. It also follows that $k = mK,\ m = 1, 2, \ldots$

$$f(x_{mK}) - f^\star \leq \frac{LD_h(X^\star, x_{(m-1)k})}{K} \leq \frac{LD_h(X^\star, x_0)}{2^{m-1}}$$

and thus (45) follows as well. □

To ease our exposition, in Proposition 6 and Proposition 7 we assumed $L_k = L$ is known and used in Step 3 of Algorithm 1. However, it is easy to see that these two results also hold if the assumption $L_k = L$ is relaxed to the assumption $L_k \leq L$ and $f(x_{k+1}) \leq \min_{y \in X}\{\langle \nabla f(x_k), y\rangle + L_k D_h(y, x_k)\}$. The latter condition is easier to implement via a standard backtracking procedure. We also assume knowledge of suitable relative smoothness constants for the choice of stepsize α_k in Step 4 of Algorithm 2 and in Step 9 of Algorithm 3 below. As in Algorithm 1, this assumption can be relaxed via a standard backtracking procedure.

5.2 Frank-Wolfe algorithm

Suppose $X \subseteq \mathbb{R}^n$ is a compact convex set and a *linear oracle* for X is available, that is, the map

$$g \mapsto \operatorname*{argmin}_{y \in X} \langle g, x \rangle$$

is computable.

The Frank-Wolfe algorithm, also known as the conditional gradient algorithm, for (2) is based on the following update for $x \in X$:

$$u := \operatorname*{argmin}_{y \in X} \langle \nabla f(x), y \rangle$$
$$x_+ := x + \alpha(u - x) \;\text{ for some }\; \alpha \in [0, 1].$$

Algorithm 2 gives a description of the Frank-Wolfe algorithm for (2).

Algorithm 2 Frank-Wolfe algorithm

1: Pick $x_0 \in X$;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad u := \operatorname{argmin}_{y \in X} \langle \nabla f(x_k), y \rangle$
4: $\quad x_{k+1} = x_k + \alpha_k(u - x_k)$ for some $\alpha_k \in [0, 1]$
5: **end for**

Let $\mathfrak{R} := \frac{\mathfrak{r}^2}{2}$ where $\mathfrak{r} : X \times X \to \mathbb{R}_+$ is the *radial* distance defined as follows: for $x, y \in X$

$$\mathfrak{r}(y, x) := \inf\{\rho > 0 : y - x = \rho \cdot (u - x) \text{ for some } u \in X\}. \tag{46}$$

Hence the relative smoothness constant $L_{f,X,\mathfrak{R}}$ is the smallest $L > 0$ such that for all $x, u \in X$ and $\alpha \in [0, 1]$

$$D_f(x + \alpha(u - x), x) \leq \frac{L\alpha^2}{2}. \tag{47}$$

Observe that the relative smoothness constant $L_{f,X,\mathfrak{R}}$ is precisely the *curvature constant* of f on X defined by Jaggi [16].

The relative quasi strong convexity constant $\mu^\star_{f,X,\mathfrak{R}}$ is the largest $\mu \geq 0$ such that for all $x \in X$

$$\frac{\mu \cdot \mathfrak{r}(\bar{x}, x)^2}{2} \leq D_f(\bar{x}, x).$$

Similarly, the relative functional growth constant $\mu^\sharp_{f,X,\mathfrak{R}}$ is the largest $\mu \geq 0$ such that for all $x \in X$

$$\frac{\mu \cdot \mathfrak{r}(\bar{x}, x)^2}{2} \leq f(x) - f^\star.$$

The next result shows the linear convergence of Algorithm 2 when $L_{f,X,\mathfrak{R}}/\mu^\star_{f,X,\mathfrak{R}}$ or $L_{f,X,\mathfrak{R}}/\mu^\sharp_{f,X,\mathfrak{R}}$ is finite. As we note below, Proposition 8 is at least as sharp as the linear convergence rates established in [3, 13].

Proposition 8. *Suppose* $L := L_{f,X,\mathfrak{R}} < \infty$ *and* $\mu := \max\{\mu^\star_{f,X,\mathfrak{R}}, \mu^\sharp_{f,X,\mathfrak{R}}/4\} > 0$. *If each stepsize* α_k *in Step 4 of Algorithm 2 is chosen via*

$$\alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]} \left\{ f(x) + \alpha \left\langle \nabla f(x), u - x \right\rangle + \frac{L\alpha^2}{2} \right\}$$

then the iterates generated by Algorithm 2 satisfy

$$f(x_k) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^\star).$$

Proof. It suffices to show that at iteration k

$$\left\langle \nabla f(x_k), x_k - u \right\rangle^2 \geq 2\mu(f(x_k) - f^\star). \tag{48}$$

Indeed, inequality (47), the choice of α_k, and (48) imply that

$$f(x_{k+1}) - f^\star \leq f(x_k) - f^\star - \frac{\left\langle \nabla f(x_k), x_k - u \right\rangle^2}{2L} \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^\star).$$

We next show (48). The construction of the radial distance and the choice of u in Algorithm 2 imply that

$$\left\langle \nabla f(x_k), x_k - \bar{x}_k \right\rangle \leq \mathfrak{r}(\bar{x}_k, x_k) \cdot \left\langle \nabla f(x_k), x_k - u \right\rangle.$$

We next consider the two possible values of $\mu = \max\{\mu^\star_{f,X,\mathfrak{R}}, \mu^\sharp_{f,X,\mathfrak{R}}/4\}$ separately.

Case 1: $\mu = \mu^\star_{f,X,\mathfrak{R}}$**.** In this case we have

$$\frac{\mu \cdot \mathfrak{r}(\bar{x}_k, x_k)^2}{2} \leq f^* - f(x_k) + \left\langle \nabla f(x_k), x_k - \bar{x}_k \right\rangle \leq f^* - f(x_k) + \mathfrak{r}(\bar{x}_k, x_k) \left\langle \nabla f(x_k), x_k - u \right\rangle.$$

Rearranging and applying the arithmetic-mean geometric-mean inequality we get

$$\left\langle \nabla f(x_k), x_k - u \right\rangle \geq \sqrt{2\mu(f(x_k) - f^\star)}.$$

Case 2: $\mu = \mu^\sharp_{f,X,\mathfrak{R}}/4$**.** In this case we have

$$2\mu \cdot \mathfrak{r}(\bar{x}_k, x_k)^2 \leq f^* - f(x_k) \leq \left\langle \nabla f(x_k), x_k - \bar{x}_k \right\rangle \leq \mathfrak{r}(\bar{x}_k, x_k) \left\langle \nabla f(x_k), x_k - u \right\rangle.$$

Therefore the last term is at least as large as the geometric mean of the first two and we get

$$\left\langle \nabla f(x_k), x_k - u \right\rangle \geq \sqrt{2\mu(f(x_k) - f^\star)}.$$

□

To conclude this subsection, we discuss some natural bounds on $L_{f,X,\mathfrak{R}}$ and $\mu^\star_{f,X,\mathfrak{R}}$. Recall that $\mathsf{ri}(X)$ denotes the relative interior of X. Similarly, let $\mathsf{rbd}(X)$ denote the relative boundary of X. As it was previously discussed in [16], from (47) it readily follows that if f is L_f-smooth on X for some norm $\|\cdot\|$ in $\mathbb{R}^n$ then

$$L_{f,X,\mathfrak{R}} \leq L_f \cdot \max_{x,y\in X} \|x-y\|^2 = L_f \cdot \mathsf{diam}(X)^2.$$

On the other hand, if f is μ_f-strongly convex on X for some norm $\|\cdot\|$ in $\mathbb{R}^n$ and the single element $x^\star \in X^\star$ satisfies $x^\star \in \mathsf{ri}(X)$ then for all $x \in X$ we have $\|x-x^\star\| = \mathfrak{r}(x^\star,x)\|x-u\| \geq \mathfrak{r}(x^\star,x)\|x^\star-u\|$ for some $u \in \mathsf{rbd}(X)$. The strong convexity of f thus implies both

$$\mu^\star_{f,X,\mathfrak{R}} \geq \mu_f \cdot \mathsf{dist}(x^\star, \mathsf{rbd}(X))^2 \;\text{ and }\; \mu^\sharp_{f,X,\mathfrak{R}} \geq \mu_f \cdot \mathsf{dist}(x^\star, \mathsf{rbd}(X))^2.$$

Therefore when f is both L_f-smooth and μ_f-strongly convex and $x^\star = \operatorname{argmin}_{x\in X} f(x) \in \mathsf{ri}(X)$ we have

$$\frac{L_{f,X,\mathfrak{R}}}{\mu^\star_{f,X,\mathfrak{R}}} \leq \frac{L_f}{\mu_f} \cdot \left(\frac{\mathsf{diam}(X)}{\mathsf{dist}(x^\star, \mathsf{rbd}(X))}\right)^2 \;\text{ and }\; \frac{L_{f,X,\mathfrak{R}}}{\mu^\sharp_{f,X,\mathfrak{R}}} \leq \frac{L_f}{\mu_f} \cdot \left(\frac{\mathsf{diam}(X)}{\mathsf{dist}(x^\star, \mathsf{rbd}(X))}\right)^2.$$

Observe that the right-hand side in both inequalities is an interesting combination of the usual condition number of f and a kind of condition number of the set X around the point $x^\star$. The first bound above and Proposition 8 yield a linear convergence result similar to [13, Theorem 2] but with a sharper rate.

The above bounds can be extended to a broader context. Suppose $f = g \circ A$ for some strongly convex function $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and $A \in \mathbb{R}^{m\times n}$. Then for all $x \in X, x^\star \in X^\star$ we have

$$\|A(x-x^\star)\| \geq \mathfrak{r}(x^\star,x) \cdot \mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X))) \geq \mathfrak{r}(\bar{x},x) \cdot \mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X))).$$

Consequently, if $X^\star \cap \mathsf{ri}(X) \neq \emptyset$ then for all $x^\star \in X^\star$

$$\mu^\star_{f,X,\mathfrak{R}} \geq \mu_g \cdot \mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X)))^2 \;\text{ and }\; \mu^\sharp_{f,X,\mathfrak{R}} \geq \mu_g \cdot \mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X)))^2.$$

Observe that $\mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X)))$ can in turn be bounded below as follows

$$\mathsf{dist}(Ax^\star, \mathsf{rbd}(A(X))) \geq \frac{1}{\|(A|\operatorname{span}(X-X))^{-1}\|} \cdot \mathsf{dist}(x^\star, \mathsf{rbd}(X)).$$

Therefore when $f = g \circ A$ where g is L_g-smooth and μ_g-strongly convex then for all $x^\star \in X^\star \cap \mathsf{ri}(X)$ both $L_{f,X,\mathfrak{R}}/\mu^\star_{f,X,\mathfrak{R}}$ and $L_{f,X,\mathfrak{R}}/\mu^\sharp_{f,X,\mathfrak{R}}$ are bounded above by

$$\frac{L_f}{\mu_f} \cdot \left(\frac{\mathsf{diam}(AX) \cdot \|(A|\operatorname{span}(X-X))^{-1}\|}{\mathsf{dist}(x^\star, \mathsf{rbd}(X))}\right)^2.$$

This bound and Proposition 8 yield a linear convergence result similar to [3, Proposition 3.2] but with a sharper rate.

5.3 Frank-Wolfe algorithm with away steps

Suppose $X \subseteq \mathbb{R}^n$ is a *polytope* and a *vertex linear oracle* for X is available, that is, the map

$$g \mapsto \operatorname*{argmin}_{y \in X} \langle g, x \rangle$$

is computable and outputs a vertex of X for all $g \in \mathbb{R}^n$.

For this kind of linear oracle, each step of the Frank-Wolfe algorithm *adds weight* to some vertex u. The basic idea of the Frank-Wolfe algorithm with away steps is to combine *regular steps* of the Frank-Wolfe algorithm with *away steps* that *reduce weight* from some vertex a. To that end, the algorithm requires an additional *vertex representation* of $x \in X$. More precisely, let $S(x) \subseteq \mathsf{vertices}(X)$ and $\lambda(x) \in \Delta(S(x)) := \{z \in \mathbb{R}_+^{S(x)} : \|z\|_1 = 1\}$ be such that

$$x = \sum_{s \in S(x)} \lambda_s(x) s \quad \text{and} \quad \lambda(x) > 0.$$

Algorithm 3 describes a Frank-Wolfe algorithm with away steps. We should highlight that although the set $\mathsf{vertices}(X)$ could be immense, the algorithm does not require it explicitly. Instead the algorithm only maintains $S(x)$ and $\lambda(x)$ that are far more manageable. Indeed, by using the IRR procedure in [2] or its modification described in [14], Step 10 in Algorithm 3 can guarantee that the sets $S(x_k)$ have size at most $n+1$ for $k = 0, 1, \ldots$.

Algorithm 3 Frank-Wolfe algorithm with away steps

1: Pick $x_0 \in \mathsf{vertices}(X)$; $S(x_0) := \{x_0\}$; $\lambda(x_0) = 1$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad u := \operatorname{argmin}_{y \in X} \langle \nabla f(x_k), y \rangle$; $a := \operatorname{argmax}_{y \in S(x_k)} \langle \nabla f(x_k), y \rangle$
4: $\quad$ **if** $\langle \nabla f(x_k), u - x_k \rangle < \langle \nabla f(x_k), x_k - a \rangle$ **then** (regular step)
5: $\quad\quad v := u - x_k$; $\alpha_{\max} = 1$;
6: $\quad$ **else** (away step)
7: $\quad\quad v := x_k - a$; $\alpha_{\max} = \frac{\lambda_a(x_k)}{1 - \lambda_a(x_k)}$;
8: $\quad$ **end if**
9: $\quad x_{k+1} := x_k + \alpha_k v$ for some $\alpha_k \in [0, \alpha_{\max}]$
10: $\quad$ update $S(x_{k+1})$ and $\lambda(x_{k+1})$
11: **end for**

Proposition 9 below establishes the linear convergence of Algorithm 3 under suitable relative smoothness and quasi strong convexity or functional growth conditions. To that end, we consider two variants of the radial distance. Let $\mathfrak{D} := \frac{\mathfrak{d}^2}{2}$ where $\mathfrak{d} : X \times X \to \mathbb{R}_+$ is the *diametral distance* defined via

$$\mathfrak{d}(y, x) := \inf\{\delta > 0 : y - x = \delta \cdot (u - w) \text{ for some } u - w \in X\}. \tag{49}$$

The relative smoothness constant $L_{f,X,\mathfrak{D}}$ is the smallest $L > 0$ such that for all $x, u, w \in X$ and $\alpha \in [0, 1]$ with $x + \alpha(u - w) \in X$

$$D_f(x + \alpha(u - w), x) \leq \frac{L\alpha^2}{2}. \tag{50}$$

The relative smoothness constant $L_{f,X,\mathfrak{D}}$ is precisely the *away curvature constant* of f on X defined by Lacoste-Julien and Jaggi [18].

To capture the appropriate relative strong convexity conditions, we rely on a more involved variant of the radial distance. For $x \in X$, let $\mathbf{S}(x)$ denote the collection of all subsets $S(x) \subseteq \mathsf{vertices}(X)$ such that x is a *positive* convex combination of the elements in $S(x)$. Let $\mathfrak{G} := \frac{\mathfrak{g}^2}{2}$ where $\mathfrak{g} : X \times X \to \mathbb{R}_+$ is defined via

$$\mathfrak{g}(y, x) := \inf \left\{ \gamma > 0 : \langle \nabla f(x), x - y \rangle \leq \gamma \cdot \min_{S(x) \in \mathbf{S}(x)} \max_{a \in S(x), u \in X} \langle \nabla f(x), a - u \rangle \right\}. \tag{51}$$

The relative strong convexity constant $\mu_{f,X,\mathfrak{G}}$ is at least as large as

$$\sup \left\{ \mu : \frac{\mu \cdot \mathfrak{g}(y, x)^2}{2} \leq D_f(y, x) \text{ for all } x, y \in X \right\}.$$

The latter quantity is precisely the *geometric strong convexity constant* defined by Lacoste-Julien and Jaggi [18, Appendix C]. Notice that it matches $\mu_{f,X,\mathfrak{G}}$ when f is strictly convex because in that case $Z_{f,X}(y) = \{y\}$ for all $y \in X$. Otherwise, it could be strictly smaller.

The relative quasi strong convexity constant $\mu^\star_{f,X,\mathfrak{G}}$ is the largest $\mu \geq 0$ such that for all $x \in X$

$$\frac{\mu \cdot \mathfrak{g}(\bar{x}, x)^2}{2} \leq D_f(\bar{x}, x).$$

Similarly, the relative functional growth constant $\mu^\sharp_{f,X,\mathfrak{G}}$ is the largest $\mu \geq 0$ such that for all $x \in X$

$$\frac{\mu \cdot \mathfrak{g}(\bar{x}, x)^2}{2} \leq f(x) - f^\star.$$

Since $\mu_{f,X,\mathfrak{G}} \leq \mu^\star_{f,X,\mathfrak{G}}$ and $\mu_{f,X,\mathfrak{G}}$ is at least as large as the geometric strong convexity constant in [18, Appendix C], the following linear convergence result is at least as sharp as the one given in [18, Theorem 8] for the Frank-Wolfe algorithm with away steps.

Proposition 9. *Suppose* $L := L_{f,X,\mathfrak{D}} < \infty$ *and* $\mu := \max\{\mu^\star_{f,X,\mathfrak{G}}, \mu^\sharp_{f,X,\mathfrak{G}}/4\} > 0$. *If each stepsize* α_k *in Step 9 of Algorithm 3 is chosen via*

$$\alpha_k = \operatorname*{argmin}_{\alpha \in [0, \alpha_{\max}]} \left\{ f(x) + \alpha \langle \nabla f(x), u - x \rangle + \frac{L\alpha^2}{2} \right\}$$

then the iterates generated by Algorithm 3 satisfy

$$f(x_k) - f^\star \leq \left(1 - \min\left\{\frac{1}{2}, \frac{\mu}{4L}\right\}\right)^{k/2} (f(x_0) - f^\star). \tag{52}$$

Proof. This proof follows a similar reasoning to the proof of Proposition 8. First we claim that at iteration k

$$\langle \nabla f(x_k), a - u \rangle^2 \geq 2\mu(f(x_k) - f^\star). \tag{53}$$

To show this claim, consider the two possible values of $\mu := \max\{\mu^\star_{f,X,\mathfrak{G}}, \mu^\sharp_{f,X,\mathfrak{G}}/4\}$ separately.

Case 1: $\mu = \mu^\star_{f,X,\mathfrak{G}}$**.** In this case we have

$$\frac{\mu \cdot \mathfrak{g}(\bar{x}_k, x_k)^2}{2} \leq f^* - f(x_k) + \langle \nabla f(x_k), x_k - \bar{x}_k \rangle \leq f^* - f(x_k) + \mathfrak{g}(\bar{x}_k, x_k) \langle \nabla f(x_k), a - u \rangle .$$

Rearranging and applying the arithmetic-mean geometric-mean inequality we get

$$\langle \nabla f(x_k), a - u \rangle \geq \sqrt{2\mu(f(x_k) - f^\star)}.$$

Case 2: $\mu = \mu^\sharp_{f,X,\mathfrak{G}}/4$**.** In this case we have

$$2\mu \cdot \mathfrak{g}(\bar{x}_k, x_k)^2 \leq f^* - f(x_k) \leq \langle \nabla f(x_k), x_k - \bar{x}_k \rangle \leq \mathfrak{g}(\bar{x}_k, x_k) \langle \nabla f(x_k), a - u \rangle .$$

Therefore the last term is at least as large as the geometric mean of the first two and we get

$$\langle \nabla f(x_k), a - u \rangle \geq \sqrt{2\mu(f(x_k) - f^\star)}.$$

To finish the proof, we next show (52) by relying on (53). To do so, we replicate some of the main ideas previously introduced in [2, 18, 28].

The choice of v at iteration k implies that

$$\langle \nabla f(x_k), v \rangle^2 \geq \frac{\langle \nabla f(x_k), a - u \rangle^2}{4} \geq \frac{\mu(f(x_k) - f^\star)}{2}. \tag{54}$$

We consider separately the three possible cases that can occur for α_k at iteration k, namely $\alpha_k < \alpha_{\max}$, $\alpha_k = \alpha_{\max} \geq 1$, and $\alpha_k = \alpha_{\max} < 1$.

Case 1: $\alpha_k < \alpha_{\max}$. In this case $|S(x_{k+1})| \leq |S(x_k)| + 1$. In addition, inequalities (50) and (54), and the choice of α_k imply that

$$f(x_{k+1}) - f(x_k) \leq -\frac{\langle \nabla f(x_k), v \rangle^2}{2L} \leq -\frac{\langle \nabla f(x_k), a - u \rangle^2}{8L} \leq -\frac{\mu}{4L}(f(x_k) - f^\star). \tag{55}$$

Case 2: $\alpha_k = \alpha_{\max} \geq 1$. In this case $|S(x_{k+1})| \leq |S(x_k)|$. In addition, inequality (50), the choice of v, and the convexity of f imply that

$$f(x_{k+1}) - f(x_k) \leq \frac{1}{2} \langle \nabla f(x_k), v \rangle \leq \frac{1}{2} \langle \nabla f(x_k), \bar{x}_k - x_k \rangle \leq -\frac{1}{2}(f(x_k) - f^\star). \tag{56}$$

Case 3: $\alpha_k = \alpha_{\max} < 1$. In this case $|S(x_{k+1})| \leq |S(x_k)| - 1$. In addition, (50) and the choice of α_k imply that

$$f(x_{k+1}) - f(x_k) \leq 0.$$

We next show that in the first k iterations Case 3 can occur at most $k/2$ times by using the argument introduced by Lacoste-Julien and Jaggi in [18]. Since $|S(x_0)| = 1$ and $|S(x_i)| \geq 1$ for $i = 1, 2, \ldots$, it follows that for each iteration when Case 3 occurred there must have been at least one previous iteration when Case 1 occurred. Hence in the first k iterations Case 3 could occur at most $k/2$ times.

To finish the proof, observe that at every iteration k when Case 1 or Case 2 occur inequalities (55) and (56) yield

$$f(x_{k+1}) - f^\star = f(x_k) - f^\star + f(x_{k+1}) - f(x_k) \leq \left(1 - \min\left\{\frac{1}{2}, \frac{\mu}{4L}\right\}\right)(f(x_k) - f^\star).$$

We note that the minimum in the last expression is is necessary because $\mu^\sharp_{f,X,\mathfrak{G}} > 2L_{f,X,\mathfrak{D}}$ may indeed occur. For a concrete example, see [28, Example 6].

□

We next discuss some bounds on $L_{f,X,\mathfrak{D}}$ and on $\mu_{f,X,\mathfrak{G}}, \mu^\star_{f,X,\mathfrak{G}}, \mu^\sharp_{f,X,\mathfrak{G}}$ in terms of the set $A := \mathsf{vertices}(X)$. We should note that the bounds below on $L_{f,X,\mathfrak{D}}$ and on $\mu_{f,X,\mathfrak{G}}$ have also been derived, albeit following a different approach, in [18, Appendix C].

From (50) it readily follows that if f is L_f-smooth on X for some norm $\|\cdot\|$ in $\mathbb{R}^n$ then

$$L_{f,X,\mathfrak{D}} \leq L_f \cdot \max_{x,y \in X} \|x - y\|^2 = L_f \cdot \mathsf{diam}(X)^2 = L_f \cdot \mathsf{diam}(A)^2.$$

On the other hand, from [28, Theorem 1] it follows that for all $x, y \in X$

$$\|y - x\| \geq \mathfrak{g}(y, x) \cdot \Phi(A)$$

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

Hence if f is μ_f-strongly convex on X for some norm $\|\cdot\|$ in $\mathbb{R}^n$ then for all $y, x \in X$ we have

$$\frac{\mu_f \Phi(A)^2 \mathfrak{g}(y, x)^2}{2} \leq \frac{\mu_f \|y - x\|^2}{2} \leq D_f(y, x)$$

and consequently

$$\mu_{f,X,\mathfrak{G}} \geq \mu_f \cdot \Phi(A)^2.$$

Therefore when f is both L_f-smooth and μ_f-strongly convex on X for some norm $\|\cdot\|$ in $\mathbb{R}^n$ we have

$$\frac{L_{f,X,\mathfrak{D}}}{\mu_{f,X,\mathfrak{G}}} \leq \frac{L_f}{\mu_f} \cdot \left(\frac{\mathsf{diam}(A)}{\Phi(A)}\right)^2.$$

Once again, the right-hand side is an interesting combination of the usual condition number of f and a kind of condition number of $A = \mathsf{vertices}(X)$. Furthermore, by

proceeding as in Example 5 it follows that when f is of the form $f(x) = \frac{1}{2}\|Bx\|_2^2$ for some $B \in \mathbb{R}^{m\times n}$ we actually have $L_{f,X,\mathfrak{D}} = \mathsf{diam}(BA)^2$ and $\mu_{f,X,\mathfrak{G}} = \Phi(BA)^2$. Thus for $f(x) = \frac{1}{2}\|Bx\|_2^2$ we have

$$\frac{L_{f,X,\mathfrak{D}}}{\mu_{f,X,\mathfrak{G}}} = \left(\frac{\mathsf{diam}(BA)}{\Phi(BA)}\right)^2.$$

This illustrates how the condition number of f relative to X depends on how the shape of X and f fit together.

We also have the following sharper lower bound on $\mu^\star_{f,X,\mathfrak{G}}$. From [28, Theorem 3] it follows that

$$\|x^\star - x\| \geq \mathfrak{g}(x^\star, x) \cdot \min_{\substack{F\in\mathsf{faces}(G)\\ \emptyset\neq F\neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A\setminus F))$$

where $G \in \mathsf{faces}(\mathsf{conv}(A))$ is the smallest face of $\mathsf{conv}(A) = X$ that contains $X^\star$. It thus follows that if f is μ_f-strongly convex on X for some norm $\|\cdot\|$ then

$$\mu^\star_{f,X,\mathfrak{G}} \geq \mu_f \cdot \min_{\substack{F\in\mathsf{faces}(G)\\ \emptyset\neq F\neq \mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A\setminus F))^2.$$

Finally we note that Theorem 4 implies that $\mu^\sharp_{f,X,\mathfrak{G}} > 0$ when f is of the form $f(x) = g(Ex) + \langle b, x\rangle$ for some strongly convex function g. Indeed, with a slight abuse of notation, let $A \in \mathbb{R}^{n\times N}$ denote the matrix whose columns are the elements of A and consider the function $\tilde{f} : \mathbb{R}^N \to \mathbb{R}$ defined via $\tilde{f} := f \circ A$. Observe that for $u, v \in \Delta_{N-1}$

$$D_f(Av, Au) = D_{\tilde{f}}(v, u) \text{ and } \mathfrak{g}(Au, Av) \leq \frac{\|u - v\|_1}{2}.$$

Consequently,

$$\mu^\sharp_{f,X,\mathfrak{G}} \geq 4\mu^\sharp_{\tilde{f},\Delta_{N-1},D}$$

for the distance function $D(v, u) := \frac{1}{2}\|v - u\|_1^2$. The functional growth constant $\mu^\sharp_{\tilde{f},\Delta_{N-1},D}$ in turn can be bounded below as detailed in Theorem 4 since $\tilde{f}$ can be written as $\tilde{f}(u) = g(EAu) + \langle b, Au\rangle$ and g is strongly convex.

The linear convergence bounds in Proposition 9 are tight modulo some small constants. This can be readily inferred from [28, Example 3 and Example 4].

References

[1] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

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

[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] S. Bubeck, Y. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[5] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[6] D. Cheung and F. Cucker. A new condition number for linear programming. *Math. Prog.*, 91(2):163–174, 2001.

[7] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar. The radius of metric regularity. *Trans. Amer. Math. Soc.*, 355(2):493–517 (electronic), 2003.

[8] D. Drusvyatskiy, M. Fazel, and S. Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

[9] M. Epelman and R. Freund. A new condition measure, preconditioners, and relations between different measures of conditioning for conic linear systems. *SIAM J. Optim.*, 12(3):627–655 (electronic), 2002.

[10] 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.

[11] R. Freund. Complexity of convex optimization using geometry-based measures and a reference point. *Math Program.*, 99:197–221, 2004.

[12] R. Freund and J. Vera. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM J. on Optim.*, 10:155–176, 1999.

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

[14] D. Gutman. Enhanced basic procedures for the projection and rescaling algorithm. *To Appear in Optimization Letters*, 2018.

[15] A. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

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

[17] S. Karimi and S. Vavasis. A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent. *arXiv preprint arXiv:1712.09498*, 2017.

[18] 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.

[19] A. Lewis. Ill-conditioned convex processes and conic linear systems. *Math. Oper. Res.*, 24(4):829–834, 1999.

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

[21] C. Ma, N. Gudapati, M. Jahani, R. Tappenden, and M. Takáč. Underestimate sequences via quadratic averaging. *arXiv preprint arXiv:1710.03695*, 2017.

[22] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

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

[24] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[25] F. Ordóñez and R. Freund. Computational experience and the explanatory value of condition measures for linear optimization. *SIAM J. on Optim.*, 14(2):307–333 (electronic), 2003.

[26] J. Peña, J. Vera, and L. Zuluaga. New characterizations of Hoffman constants for system of linear constraints. *arXiv preprint arXiv:1905.20894*, 2019.

[27] J. Peña. Understanding the geometry on infeasible perturbations of a conic linear system. *SIAM J. on Optim.*, 10:534–550, 2000.

[28] J. Peña and D. Rodríguez. Polytope conditioning and linear convergence of the Frank-Wolfe algorithm. *To Appear in Mathematics of Operations Research*, 2018.

[29] A. Ramdas and J. Peña. Towards a deeper geometric, analytic and algorithmic understanding of margins. *Optimization Methods and Software*, 31(2):377–391, 2016.

[30] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM J. on Optim.*, 5:506–524, 1995.

[31] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Math. Programming*, 70(3, Ser. A):279–351, 1995.

[32] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, pages 1–30, 2018.

A Proof of Proposition 5

The construction of $T_X(x;A,S)$ implies $T_X(x;A,S) \subseteq T_X(x)$ and $\|(A|T_X(x;A,S))^{-1}\| \leq \|(A|T_X(x))^{-1}\|$ for all $x \in X$. Hence

$$\sup_{C\in\mathcal{T}(A|X,S)} \|(A|C)^{-1}\| \leq \max_{C\in\mathcal{T}(X)} \|(A|C)^{-1}\| = \max_{C\in\mathcal{T}(A|X)} \|(A|C)^{-1}\|$$

where the last step follows from [26, Lemma 1]. This proves the second inequality in (33).

Let $H := \sup_{C\in\mathcal{T}(A|X,S)} \|(A|C)^{-1}\|$. The first inequality in (33) can be stated as follows: for all $y \in S$ and $x \in X$

$$\|Z_{A,X}(y) - x\| \leq H \cdot \|Ay - Ax\|. \tag{57}$$

We prove (57) by contradiction. Suppose that there exist $y \in S$ and $x \in X \setminus Z_{A,X}(y)$ such that $\|Z_{A,X}(y) - x\| > H \cdot \|Ay - Ax\|$. That is,

$$A\tilde{y} = Ay, y \in X \Rightarrow \|\tilde{y} - x\| > H \cdot \|Ay - Ax\|. \tag{58}$$

Let $v := (Ay - Ax)/\|Ay - Ax\|$ and consider the convex optimization problem

$$\begin{array}{ll} \max\limits_{u,t} & t \\ & Au = tv \\ & x + u \in X \\ & \|u\| \leq H \cdot t. \end{array} \tag{59}$$

Observe that $v \in A(T_X(x;A,S))$ since $y - x \in T_X(x;A,S)$. Thus there exists $u \in T_X(x;A,S)$ such that $Au = v$ and

$$\|u\| \leq \|(A|T_X(x;A,S))^{-1}\| \leq H.$$

Therefore there exists (u,t) feasible for (59) with $t > 0$. On the other hand, (58) implies that there does not exist any (u,t) feasible for (59) with $t = \|Ay - Ax\|$. It thus follows that (59) has an optimal solution $(\hat{u},\hat{t})$ with $0 < \hat{t} < \|Ay - Ax\|$. Now consider the modification of (59) obtained by replacing x with $x + \hat{u} \in X$:

$$\begin{array}{ll} \max\limits_{u,t} & t \\ & Au = tv \\ & x + \hat{u} + u \in X \\ & \|u\| \leq H \cdot t. \end{array} \tag{60}$$

Proceeding as above with $x+\hat{u}$ in lieu of x it follows that (60) has an optimal solution (u',t') with $0<t'<\|Ay-Ax\|-\hat{t}$. In particular, $(\hat{u}+u',\hat{t}+t')$ is a feasible solution to (59) with $\hat{t}+t'>\hat{t}$ which contradicts the optimality of $(\hat{u},\hat{t})$. We therefore conclude that (57) must hold and thus (34) is proven.

We next prove (34) when $A(S)$ is convex. To that end, suppose $C\in\mathcal{T}(A|X,S)$ and $0<\epsilon<\|(A|C)^{-1}\|$. Then $C=T_X(\hat{x};A,S)$ for some $\hat{x}\in X$. Let $\hat{v}\in C$ be such that $A\hat{v}\neq 0$ and $\|v\|\geq(\|(A|C)^{-1}\|-\epsilon)\cdot\|A\hat{v}\|$ for all $v\in C$ with $Av=A\hat{v}$. By scaling $\hat{v}$ if necessary we can assume that $A(\hat{x}+\hat{v})\in\mathsf{conv}(A(S))=A(S)$ and thus $A(\hat{x}+\hat{v})=A\hat{y}$ for some $\hat{y}\in S$. Observe that $\hat{x}+v\in Z_{A,X}(\hat{y})$ implies both $v\in C$ and $Av=A\hat{v}$. It thus follows that

$$\frac{1}{\|(A|C)^{-1}\|-\epsilon}\geq\frac{\|A\hat{v}\|}{\|Z_{A,X}(\hat{y})-\hat{x}\|}=\frac{\|A\hat{y}-A\hat{x}\|}{\|Z_{A,X}(\hat{y})-\hat{x}\|}\geq\inf_{\substack{y\in S,\,x\in X\\ x\notin Z_{A,X}(y)}}\frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}.$$

Since this holds for all $C\in\mathcal{T}(A|X,S)$ and $0<\epsilon<\|(A|C)^{-1}\|$ identity (34) follows. $\square$