Affine invariant convergence rates of the conditional gradient method

Javier F. Peña*

December 13, 2021

Abstract

We show that the conditional gradient method for the convex composite problem

$$\min_x \{f(x) + \Psi(x)\}$$

generates primal and dual iterates with a duality gap converging to zero provided a suitable *growth property* holds and the algorithm makes a judicious choice of stepsizes. The rate of convergence of the duality gap to zero ranges from sublinear to linear depending on the degree of the growth property. The growth property and convergence results depend exclusively on the pair (f, Ψ). They are both affine invariant and norm-independent.

1 Introduction

We consider the conditional gradient method for the composite minimization problem

$$\min_x \{f(x) + \Psi(x)\}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are closed convex functions, f is differentiable, and there is an available oracle for the mapping

$$g \mapsto \arg\min\{\langle g, y\rangle + \Psi(y)\}.$$

The traditional format of the conditional gradient method (also known as the Frank-Wolfe algorithm) as discussed in [3, 5] corresponds to the case when Ψ is the indicator function of a compact convex set in $\mathbb{R}^n$. Some of the most attractive features of the conditional gradient method are its affine invariance, norm-independence, and lack of reliance on any projection mapping. These features stand in stark contrast to those of most other first-order methods.

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

We show that the conditional gradient method generates primal and dual iterates with a duality gap converging to zero provided a suitable *growth property* holds (see (8)) and the algorithm makes a judicious choice of stepsizes (see (7) and (15)). The rate of convergence of the duality gap to zero ranges from sublinear to linear (see Theorem 1 and Proposition 2) depending on the degree of the growth property. The growth property and convergence results depend exclusively on the pair (f, Ψ). They are both affine invariant and norm-independent. In fact, our main developments do not rely on any norms at all.

Our work is inspired by recent developments of Nesterov [11], Ghadimi [4], and Kerdreux et al [7]. Under the assumption that the domain of Ψ is bounded, Nesterov [11] established a $\mathcal{O}(1/k^\nu)$ convergence rate for the conditional gradient method when ∇f is ν-Hölder continuous. Nesterov [11] also showed the stronger convergence rate $\mathcal{O}(1/k^{2\nu})$ under the additional assumption that Ψ is strongly convex. Ghadimi [4] showed that under the latter assumptions a stronger linear rate of convergence is achievable for a judicious choice of stepsizes. On the other hand, Kerdreux et al. [7] considered the case when ∇f is Lipschitz continuous and Ψ is the indicator function of a uniformly convex set. Kerdreux et al. [7] showed that the convergence rate ranges from sublinear to linear depending on the degree of uniform convexity of the set. The results in [4, 7, 11] are all norm-dependent since they rely on Hölder continuity and uniform convexity. The norm-dependence of the results in [4, 7, 11] in turn implies that they are also affine dependent when the norms are affine dependent, which is typically the case.

The affine invariance and norm-independence properties of the conditional gradient method suggest that there should be an affine invariant and norm-independent approach to obtain the sublinear to linear spectrum of convergence rates. The central contribution of this paper is to show that this is indeed the case. To that end, we establish general convergence results (Theorem 1 and Proposition 2) provided a *growth property* (8) holds and the stepsizes are judiciously chosen. The growth property (8) can be seen as an extension of the *curvature constant* proposed by Jaggi [5]. It also has a flavor similar to that of the *relative smoothness* and *relative continuity* properties in the context of Bregman proximal methods as discussed in [1, 9, 10, 12–14]. We show that our main convergence results (Theorem 1 and Proposition 2) subsume some of the norm-dependent convergence results in [4,7,11]. Our convergence results are stated in terms of the duality gap between the primal and dual iterates that the algorithm generates. The convergence of the duality gap automatically implies the convergence of the suboptimality gap, as in the statements in [7, 11].

Our developments, driven by affine-invariance, are in the same spirit as those of Kerdreux et al. [8]. However, the approach in [8] applies only to the case when Ψ is the indicator function of a strongly convex set and focuses on an affine-invariant *directional smoothness* property that ensures linear convergence. By contrast, our work applies to general Ψ and relies on the growth property (8) that leads to convergence rates covering the entire range from sublinear to linear.

The remaining sections of the paper is organized as follows. Section 2 presents our main developments, namely the affine invariant and norm-independent growth prop-

erty (8) and convergence results for the conditional gradient method. For ease of exposition, we include two convergence results. First, Theorem 1 relies on the assumption that the stepsizes are chosen via the exact line-search (7). Proposition 2 shows that the convergence rates in Theorem 1 continue to hold, albeit with some more complicated constants, for the more flexible and easily implementable choice of stepsizes (15). Section 3 details some classes of problems that satisfy the growth property (8). As a byproduct, we recover the norm-dependent convergence rates established in [4, 7, 11]. Section 4 presents a more flexible and general *local* growth property that in turn yields local convergence results for the conditional gradient method.

2 Growth property and affine invariant convergence

Consider the composite minimization problem

$$\min_x \{f(x) + \Psi(x)\} \tag{1}$$

where both $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are closed convex functions. Algorithm 1 describes the conditional gradient algorithm for (1). The algorithm relies on the following assumptions about (1):

A1. The function f is differentiable on $\mathrm{dom}(\Psi)$.

A2. A point in the following set of minimizers exists and is computable for all $x \in \mathrm{dom}(\Psi)$

$$\arg\min_s \{\langle \nabla f(x), s\rangle + \Psi(s)\}.$$

Our main developments (Theorem 1 and Proposition 2) concern the behavior of the gap between the *primal* iterates $x_k,\ k = 0, 1, \ldots$ generated by Algorithm 1 and the *dual* iterates $g_k := \nabla f(x_k),\ k = 0, 1, \ldots$ that Algorithm 1 also generates for the Fenchel dual of (1), namely

$$\max_u \{-f^*(u) - \Psi^*(-u)\}.$$

Here f^* and Ψ^* denote the conjugates of f and Ψ respectively. We will rely on the observation that for $g \in \mathbb{R}^n$, the set $\arg\min_y \{\langle g, y\rangle + \Psi(y)\}$ is precisely the subdifferential $\partial\Psi^*(-g)$. We will also rely on the observation that the closedness of f and Ψ together with Assumptions A1 and A2 imply that $\mathrm{dom}(\Psi) \subseteq \mathrm{dom}(f)$ and $\mathrm{dom}(f^*) \subseteq -\mathrm{dom}(\Psi^*)$.

Algorithm 1 Conditional gradient algorithm

input: (f, Ψ), $x_0 \in \mathrm{dom}(\Psi)$
for $k = 0, 1, 2, \ldots$ **do**
  pick $s_k \in \arg\min_y \{\langle \nabla f(x_k), y\rangle + \Psi(y)\}$ and $\theta_k \in [0, 1]$
  let $x_{k+1} := (1 - \theta_k)x_k + \theta_k s_k$
end for

Starting from a current iterate x, Algorithm 1 updates x to $x_+ := x + \theta(s - x)$ where $s \in \partial\Psi^*(-g)$ for $g = \nabla f(x)$, and $\theta \in [0, 1]$. The choice of $g = \nabla f(x)$ and $s \in \partial\Psi^*(-g)$ yields the identity

$$f(x) + f^*(g) + \Psi(s) + \Psi^*(-g) = \langle g, x - s\rangle\,,$$

which can be equivalently stated as the *Wolfe gap:*

$$f(x) + \Psi(x) + f^*(g) + \Psi^*(-g) = \langle g, x - s\rangle + \Psi(x) - \Psi(s) \tag{2}$$

or as

$$f(s) + \Psi(s) + f^*(g) + \Psi^*(-g) = f(s) - f(x) - \langle g, s - x\rangle = D_f(s, x), \tag{3}$$

where D_f denotes the *Bregman distance* of f, that is, the function

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

Our main developments are stated in terms of the functions gap and $\mathcal{D}$ described next. Define the duality gap function $\mathsf{gap} : \mathrm{dom}(\Psi) \times \mathrm{dom}(f^*) \to \mathbb{R}$ as follows

$$\mathsf{gap}(x, u) := f(x) + \Psi(x) + f^*(u) + \Psi^*(-u). \tag{4}$$

Define $\mathcal{D} : \mathrm{dom}(\Psi) \times \mathrm{dom}(\Psi) \times [0, 1] \to \mathbb{R}$ as follows

$$\mathcal{D}(x, s, \theta) := D_f(x + \theta(s - x), x) + \Psi(x + \theta(s - x)) - (1 - \theta)\Psi(x) - \theta\Psi(s). \tag{5}$$

The following proposition establishes a fundamental identity connecting gap and $\mathcal{D}$.

Proposition 1. *Suppose $x \in \mathrm{dom}(\Psi), g = \nabla f(x)$, and $s \in \partial\Psi^*(-g)$ Then for $\theta \in [0, 1]$ we have*

$$\mathsf{gap}(x + \theta(s - x), g) = (1 - \theta)\mathsf{gap}(x, g) + \mathcal{D}(x, s, \theta). \tag{6}$$

Proof. Adding (2) times $(1 - \theta)$ and (3) times θ we get

$$(1 - \theta)(f + \Psi)(x) + \theta(f + \Psi)(s) + f^*(g) + \Psi^*(-g) = (1 - \theta)\mathsf{gap}(x, g) + \theta D_f(s, x)$$

Thus

$$\begin{aligned}
&\mathsf{gap}(x + \theta(s - x), g)\\
&= (f + \Psi)(x + \theta(s - x)) + f^*(g) + \Psi^*(-g)\\
&= (1 - \theta)\mathsf{gap}(x, g) + \theta D_f(s, x) + (f + \Psi)(x + \theta(s - x)) - (1 - \theta)(f + \Psi)(x) - \theta(f + \Psi)(s).
\end{aligned}$$

To finish, observe that

$$\begin{aligned}
&\theta D_f(s, x) + (f + \Psi)(x + \theta(s - x)) - (1 - \theta)(f + \Psi)(x) - \theta(f + \Psi)(s)\\
&= D_f(x + \theta(s - x), x) + \Psi(x + \theta(s - x)) - (1 - \theta)\Psi(x) - \theta\Psi(s)\\
&= \mathcal{D}(x, s, \theta).
\end{aligned}$$

□

Theorem 1 below shows that Algorithm 1 generates primal and dual iterates whose duality gap converges to zero provided the growth property (8) below holds and Algorithm 1 makes a judicious choice of stepsizes at each iteration. For a cleaner statement, Theorem 1 assumes that Algorithm 1 chooses the stepsize θ_k via the following line-search procedure

$$\theta_k := \operatorname*{arg\,min}_{\theta\in[0,1]} \left\{(1-\theta)\mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta)\right\}. \tag{7}$$

As Proposition 2 shows, the main conclusions in Theorem 1 continue to hold, albeit with more complicated constants, for a more flexible procedure to choose the stepsize.

The gap reduction identity (6) in Proposition 1 implies that (7) can be equivalently stated as

$$\theta_k = \operatorname*{arg\,min}_{\theta\in[0,1]} (f+\Psi)(x_k + \theta(s_k - x_k)).$$

We prefer (7) because it directly connects with the gap reduction identity (6).

Theorem 1 hinges on the following growth property connecting the functions $\mathcal{D}$ and gap. Suppose $q > 1$ and $r \in [0, 1]$. We shall say that the pair $(\mathcal{D}, \mathsf{gap})$ satisfies the (q, r)-*growth property* if there exist a finite constant $M > 0$ such that for all $x \in \mathrm{dom}(\Psi)$, $g := \nabla f(x)$, and $s \in \partial\Psi^*(-g)$

$$\mathcal{D}(x, s, \theta) \le \frac{M\theta^q}{q} \cdot \mathsf{gap}(x, g)^r \text{ for all } \theta \in [0, 1]. \tag{8}$$

To ease our exposition, we shall assume without loss of generality that $M \ge 1$ whenever (8) holds. The (q, r)-growth property (8) can be seen as a generalization the *curvature constant* defined by Jaggi [5]. Indeed, suppose $\Psi = \delta_C$ for some some compact convex set $C \subseteq \mathbb{R}^n$, and $q = 2$ and $r = 0$. In this case the (q, r)-growth property (8) can be rewritten as

$$D_f(x + \theta(s - x), x) \le \frac{M\theta^2}{2} \text{ for all } \theta \in [0, 1]. \tag{9}$$

The smallest constant M that satisfies (9) is precisely Jaggi's curvature constant. The growth property (8) has a flavor similar to that of the *relative smoothness* and *relative continuity* properties in the context of Bregman proximal methods as discussed in [1, 9, 10, 12–14].

The crux of Theorem 1 is to combine the gap reduction identity (6), the growth property (8), and the following technical lemma due to Borwein, Li, and Yao [2, Lemma 4.1].

Lemma 1. *Suppose $p > 0$ and $\beta_k, \delta_k \ge 0$ are such that $\beta_{k+1} \le \beta_k(1 - \delta_k\beta_k^p)$ for $k = 0, 1, \ldots$. Then*

$$\beta_k \le \left(\beta_0^{-p} + p \cdot \sum_{i=0}^{k-1} \delta_i\right)^{-\frac{1}{p}}.$$

Our two central results, namely Theorem 1 and Proposition 2, concern the sequence gap_k, $k = 0, 1, \ldots$ of *best duality gaps* defined as follows. For $k = 0, 1, \ldots$ let

$$\mathsf{gap}_k := \min_{i=0,1,\ldots,k} \mathsf{gap}(x_k, g_i) \tag{10}$$

where $g_i = \nabla f(x_i)$ for $i = 0, 1, \ldots$.

Theorem 1. *Suppose Algorithm 1 chooses $\theta_k \in [0,1]$ via (7) and $q > 1$ and $r \in [0,1]$ are such that $(\mathcal{D}, \mathsf{gap})$ satisfy the (q,r)-growth property (8) for some finite $M > 0$. Then for $k = 0, 1, \ldots$*

$$\mathsf{gap}_{k+1} \leq \begin{cases} \mathsf{gap}_k \left(1 - \frac{q-1}{q} \cdot \left(\frac{\mathsf{gap}_k^{1-r}}{M}\right)^{\frac{1}{q-1}}\right) & \text{if } \mathsf{gap}_k^{1-r} \geq M \\ \mathsf{gap}_k \left(1 - \frac{q-1}{q}\right) & \text{otherwise.} \end{cases} \tag{11}$$

When $r = 1$ we have linear convergence

$$\mathsf{gap}_k \leq \mathsf{gap}_0 \left(1 - \frac{q-1}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}}\right)^k. \tag{12}$$

When $r \in [0,1)$ we have an initial linear convergence regime

$$\mathsf{gap}_k \leq \mathsf{gap}_0 \left(1 - \frac{q-1}{q}\right)^k, \; k = 0, 1, 2, \ldots, k_0 \tag{13}$$

where k_0 is the smallest k such that $\mathsf{gap}_k^{1-r} \leq M$. Then for $k \geq k_0$ we have a sublinear convergence regime

$$\mathsf{gap}_k \leq \left(\mathsf{gap}_{k_0}^{\frac{r-1}{q-1}} + \frac{1-r}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}} \cdot (k - k_0)\right)^{\frac{q-1}{r-1}}. \tag{14}$$

Proof. Identity (6) in Proposition 1 and equations (7) and (8) imply that

$$\begin{aligned} \mathsf{gap}(x_{k+1}, g_k) &= \min_{\theta \in [0,1]} \{(1-\theta)\mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta)\} \\ &\leq \mathsf{gap}(x_k, g_k) + \min_{\theta \in [0,1]} \left\{-\theta \cdot \mathsf{gap}(x_k, g_k) + \frac{M\theta^q}{q} \cdot \mathsf{gap}(x_k, s_k)^r\right\}. \end{aligned}$$

Thus (10) implies that

$$\begin{aligned} \mathsf{gap}_{k+1} &\leq \mathsf{gap}_k + \min_{\theta \in [0,1]} \left\{-\theta \cdot \mathsf{gap}_k + \frac{M\theta^q}{q} \cdot \mathsf{gap}_k^r\right\} \\ &\leq \begin{cases} \mathsf{gap}_k \left(1 - \frac{q-1}{q} \cdot \left(\frac{\mathsf{gap}_k^{1-r}}{M}\right)^{\frac{1}{q-1}}\right) & \text{if } \mathsf{gap}_k^{1-r} \leq M\theta_{\max}^{q-1} \\ \mathsf{gap}_k \left(1 - \frac{q-1}{q}\right) & \text{otherwise.} \end{cases} \end{aligned}$$

The second step above follows because the minimum is attained at

$$\hat{\theta} = \min\left\{\left(\frac{\mathsf{gap}_k^{1-r}}{M}\right)^{\frac{1}{q-1}}, 1\right\}$$

and in both possible cases we have $M\hat{\theta}^{q-1} \leq \mathsf{gap}_k^{1-r}$ so

$$\mathsf{gap}_k - \hat{\theta} \cdot \mathsf{gap}_k + \frac{M\hat{\theta}^q}{q} \cdot \mathsf{gap}_k^r \leq \mathsf{gap}_k \cdot \left(1 - \frac{q-1}{q} \cdot \hat{\theta}\right).$$

Therefore (11) is established. Inequality (11) automatically implies (12) when $r = 1$ and also (13) for $k \leq k_0$ when $r \in [0, 1)$. When $r \in [0, 1)$ and $k \geq k_0$ inequality (14) follows from (11) and Lemma 1 applied to $p := \frac{1-r}{q-1}$ and $\beta_k := \mathsf{gap}_k, \delta_k := \frac{q-1}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}}$ for $k = k_0, k_0 + 1, \ldots$. □

Theorem 1 shows that for fixed $q > 1$, the value of $r \in [0, 1]$ in the (q, r)-growth property (8) yields a convergence rate for $\mathsf{gap}_k \to 0$ somewhere in the spectrum from the sublinear rate $\mathcal{O}(k^{1-q})$ when $r = 0$ to a linear rate when $r = 1$. The convergence rate increases as r increases to 1. For fixed $r \in (0, 1)$ the converge rate also increases as the parameter $q > 1$ in the (q, r)-growth property (8) increases.

We next describe a more flexible alternative to (7) to choose the stepsize θ_k. Suppose $c, \rho \in (0, 1)$. A simple backtracking procedure can easily choose $\theta_k \in [0, 1]$ such that $\rho \cdot \hat{\theta}_k \leq \theta_k \leq \hat{\theta}_k$ for

$$\hat{\theta}_k := \max\left\{\theta \in [0, 1] : (1 - \theta)\mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta) \leq (1 - c \cdot \theta)\mathsf{gap}(x_k, g_k)\right\}. \quad (15)$$

A straightforward modification of the proof of Theorem 1 yields the following result.

Proposition 2. *Suppose $c, \rho \in (0, 1)$ are such that $c + \rho > 1$ and Algorithm 1 chooses $\theta_k \in [0, 1]$ so that $\rho \cdot \hat{\theta}_k \leq \theta_k \leq \hat{\theta}_k$ for $\hat{\theta}_k$ as in (15). Suppose $q > 1$ and $r \in [0, 1]$ are such that $(\mathcal{D}, \mathsf{gap})$ satisfy the (q, r)-growth property (8) for some finite $M > 0$. Then for $k = 0, 1, \ldots$*

$$\mathsf{gap}_{k+1} \leq \begin{cases} \mathsf{gap}_k \left(1 - (c + \rho - 1) \cdot \left(\frac{q(1-c)}{M} \cdot \mathsf{gap}_k^{1-r}\right)^{\frac{1}{q-1}}\right) & \text{if } \mathsf{gap}_k^{1-r} \leq \frac{M}{q(1-c)} \\ \mathsf{gap}_k \cdot (1 - (c + \rho - 1)) & \text{otherwise.} \end{cases} \quad (16)$$

When $r = 1$ we have linear convergence

$$\mathsf{gap}_k \leq \mathsf{gap}_0 \left(1 - (c + \rho - 1) \cdot \left(\frac{q(1-c)}{M}\right)^{\frac{1}{q-1}}\right)^k. \quad (17)$$

When $r \in [0, 1)$ we have an initial linear convergence regime

$$\mathsf{gap}_k \leq \mathsf{gap}_0 \left(1 - (c + \rho - 1)\right)^k, \; k = 0, 1, 2, \ldots, k_0 \quad (18)$$

where k_0 is the smallest k such that $\mathsf{gap}_k^{1-r} \leq \frac{M}{q(1-c)}$. Subsequently for $k \geq k_0$ we have a sublinear convergence regime

$$\mathsf{gap}_k \leq \left(\mathsf{gap}_{k_0}^{\frac{r-1}{q-1}} + \frac{(1-r)(c+\rho-1)}{q-1} \cdot \left(\frac{q(1-c)}{M}\right)^{\frac{1}{q-1}} \cdot (k - k_0)\right)^{\frac{q-1}{r-1}}. \quad (19)$$

Proof. As noted above, this is a modification of the proof of Theorem 1. Identity (6) in Proposition 1 yields

$$\mathsf{gap}(x_{k+1}, g_k) = (1-\theta_k)\mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta_k).$$

Hence (10) implies that

$$\mathsf{gap}_{k+1} \leq \mathsf{gap}_k - \theta_k \cdot \mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta_k).$$

Since $\rho \cdot \hat{\theta}_k \leq \theta_k \leq \hat{\theta}_k$ for $\hat{\theta}_k$ as in (15), it follows that

$$\begin{aligned}
\mathsf{gap}_{k+1} &\leq \mathsf{gap}_k - \rho \cdot \hat{\theta}_k \cdot \mathsf{gap}(x_k, g_k) + \mathcal{D}(x_k, s_k, \theta_k) \\
&\leq \mathsf{gap}_k - \rho \cdot \hat{\theta}_k \cdot \mathsf{gap}(x_k, g_k) + (1-c) \cdot \theta_k \cdot \mathsf{gap}(x_k, g_k) \\
&\leq \mathsf{gap}_k - (c+\rho-1) \cdot \hat{\theta}_k \cdot \mathsf{gap}(x_k, g_k) \\
&\leq \mathsf{gap}_k (1 - (c+\rho-1) \cdot \hat{\theta}_k). \qquad (20)
\end{aligned}$$

The growth property (8) and (10) imply that the minimizer $\hat{\theta}_k$ in (15) satisfies

$$\begin{aligned}
\hat{\theta}_k &= \max\{\theta \in [0,1] : \mathcal{D}(x_k, s_k, \theta) \leq (1-c) \cdot \theta \cdot \mathsf{gap}(x_k, g_k)\} \\
&\geq \max\left\{\theta \in [0,1] : \frac{M\theta^q}{q} \mathsf{gap}(x_k, g_k)^r \leq (1-c) \cdot \theta \cdot \mathsf{gap}(x_k, g_k)\right\} \\
&= \min\left\{\left(\frac{q(1-c)}{M} \cdot \mathsf{gap}(x_k, g_k)^{1-r}\right)^{\frac{1}{q-1}}, 1\right\} \\
&\geq \min\left\{\left(\frac{q(1-c)}{M} \cdot \mathsf{gap}_k^{1-r}\right)^{\frac{1}{q-1}}, 1\right\}. \qquad (21)
\end{aligned}$$

Inequality (16) follows by putting together (20) and (21). Inequality (16) automatically implies (17) when $r = 1$ and also (18) for $k \leq k_0$ when $r \in [0,1)$. When $r \in [0,1)$ and $k \geq k_0$ inequality (19) follows from (16) and Lemma 1 applied to $p := \frac{1-r}{q-1}$ and $\beta_k := \mathsf{gap}_k, \delta_k := (c+\rho-1)\left(\frac{q(1-c)}{M}\right)^{\frac{1}{q-1}}$ for $k = k_0, k_0+1, \ldots$. □

We note that expression (16) in Proposition 2 is identical to expression (11) in Theorem 1 for the ideal choice $\rho = 1$ and $c = (q-1)/q$. Likewise for (17), (19) and (12), (14).

To conclude this section, we next discuss the norm-independence and affine invariance of our developments. The growth property (8) as well as Theorem 1 and Proposition 2 are automatically norm-independent since they do not rely on any norms. They are also affine invariant as we next detail.

Consider an affine *reparametrization* of $\mathbb{R}^n$ of the form $x := A\tilde{x} + b$ where $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear bijection. Let $\tilde{f}, \tilde{\Psi} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be defined as the corresponding reparametrizations of f and Ψ, namely

$$\tilde{f}(\tilde{x}) := f(A\tilde{x} + b) \quad \text{and} \quad \tilde{\Psi}(\tilde{x}) := \Psi(A\tilde{x} + b).$$

Let $A^* : \mathbb{R}^n \to \mathbb{R}^n$ denote the adjoint of A. Some straightforward calculations show that if $x = A\tilde{x} + b$ then

$$\nabla \tilde{f}(\tilde{x}) = A^* \nabla f(x) \text{ and } \tilde{f}^*(A^* u) = f^*(u) - \langle b, u \rangle \text{ for all } u \in \mathbb{R}^n,$$

and likewise for Ψ in lieu of f. Consequently, if $x = A\tilde{x} + b$ then

$$\partial \Psi^*(-A^* \nabla f(x)) = A \partial \tilde{\Psi}^*(-\nabla \tilde{f}(\tilde{x})) + b.$$

Hence if $x = A\tilde{x} + b$ then for $g := \nabla f(x)$ and $\tilde{g} := \nabla \tilde{f}(\tilde{x})$ we have $\tilde{g} = A^* g$ and $\partial \Psi^*(-g) = A \partial \Psi^*(-\tilde{g}) + b$. Thus we can assume that the oracle in Assumption A2 yield respectively $s \in \partial \Psi^*(-g)$ and $\tilde{s} \in \partial \tilde{\Psi}^*(-\tilde{g})$ that satisfy $s = A\tilde{s} + b$. The identities $x = A\tilde{x} + b,\ \tilde{g} = A^* g,\ s = A\tilde{s} + b$ in turn imply that

$$\mathsf{gap}(x, g) = \widetilde{\mathsf{gap}}(\tilde{x}, \tilde{g})$$

and

$$\mathcal{D}(x, s, \theta) = \widetilde{\mathcal{D}}(\tilde{x}, \tilde{s}, \theta) \text{ for all } \theta \in [0, 1].$$

It thus follows that the growth property (8), stepsize selection procedures (7) and (15), as well as Theorem 1 and Proposition 2 are all affine invariant.

3 Growth property for some classes of problems

Examples 1 through 3 below detail some classes of problems that satisfy the growth property (8). As we explain below, Examples 1 through 3 together with Theorem 1 yield the same kind of convergence rates in some of the results in [4, 7, 11]. These examples rely on the concepts of uniform smoothness and uniform convexity that we recall next. A detailed discussion on these concepts is presented in [6]. Both uniform smoothness and uniform convexity are stated and formalized in terms of some norm. Thus throughout this section assume that $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$ and let $\|\cdot\|^*$ denote its dual norm. It is worth highlighting that this section is the only portion of the paper that relies on norms.

A convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is *uniformly smooth* on $C \subseteq \mathrm{dom}(f)$ with respect to $\|\cdot\|$ if there exist constants $q \in (1, 2]$ and $L > 0$ such that for all $x, y \in C$ and $\theta \in [0, 1]$

$$f(x + \theta(y - x)) \geq (1 - \theta) f(x) + \theta f(y) - \frac{L}{q} \theta (1 - \theta) \|y - x\|^q. \tag{22}$$

It is easy to see that (22) holds for $q = \nu + 1$ when ∇f is ν-Hölder continuous on C.

As detailed in [6], when f is differentiable, property (22) implies that for all $x, y \in C$

$$D_f(y, x) \leq \frac{L}{q} \|y - x\|^q.$$

In particular, if (22) holds on $C := \mathrm{dom}(\Psi)$ then for all $x, y \in \mathrm{dom}(\Psi)$ and $\theta \in [0,1]$ the function $\mathcal{D}$ defined via (5) satisfies

$$\mathcal{D}(x, s, \theta) \leq D_f(x + \theta(s - x), x) \leq \frac{L}{q}\theta^q \|s - x\|^q. \tag{23}$$

A function $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is *uniformly convex* on $C \subseteq \mathrm{dom}(\Psi)$ with respect to $\|\cdot\|$ if there exist constants $p \geq 2$ and $\mu > 0$ such that for all $x, y \in C$ and $\theta \in [0,1]$

$$\Psi(x + \theta(y - x)) \leq (1 - \theta)\Psi(x) + \theta\Psi(y) - \frac{\mu}{p}\theta(1 - \theta)\|y - x\|^p. \tag{24}$$

Again, as detailed in [6], property (24) implies that for all $x, y \in C$ and $g \in \partial\Psi(y)$

$$\Psi(x) \geq \Psi(y) + \langle g, s - y\rangle + \frac{\mu}{p}\|x - y\|^p.$$

In particular, property (24) implies that if $C = \mathrm{dom}(\Psi)$ and $s = \arg\min_y \Psi(y)$ then for all $x \in \mathrm{dom}(\Psi)$

$$\Psi(x) - \Psi(s) \geq \frac{\mu}{p}\|x - s\|^p. \tag{25}$$

A closed convex set $C \subseteq \mathbb{R}^n$ is *uniformly convex* if there exist constants $p \geq 2$ and $c > 0$ such that for all $x, y \in C$, $\theta \in [0,1]$, and $z \in \mathbb{R}^n$ with $\|z\| \leq 1$

$$x + \theta(y - x) + \frac{\mu}{p}\theta(1 - \theta)\|y - x\|^p z \in C. \tag{26}$$

Similar to the above implications (22) $\Rightarrow$ (23) and (24) $\Rightarrow$ (25), property (26) implies that if $g \in \mathbb{R}^n$ and $s = \arg\min_{y \in C} \langle g, y\rangle$ then for all $x \in C$

$$\langle g, x - s\rangle \geq \frac{\mu}{p}\|g\|^* \cdot \|x - s\|^p. \tag{27}$$

Example 1. *The pair* $(\mathcal{D}, \mathsf{gap})$ *satisfies the* (q, r)*-growth property* (8) *for* q *and* $r = 0$ *if* f *is* q*-uniformly smooth on* $\mathrm{dom}(\Psi)$ *and* $\mathrm{dom}(\Psi)$ *bounded.*
Detail. Suppose $q \in (1, 2]$ is such that (22) holds for some $L > 0$. Then (23) implies that (8) holds for q and $r = 0$ with $M = L \cdot \mathsf{diam}(\mathrm{dom}(\Psi))^q$ where

$$\mathsf{diam}(\mathrm{dom}(\Psi)) = \max\{\|x - y\| : x, y \in \mathrm{dom}(\Psi)\} < \infty.$$

□

Observe that f is $(\nu + 1)$-uniformly smooth on $\mathrm{dom}(\Psi)$ if ∇f is ν-Hölder continuous on $\mathrm{dom}(\Psi)$. Thus Example 1 and Theorem 1 show that

$$\mathsf{gap}_k = \mathcal{O}(k^{-\nu})$$

when ∇f is ν-Hölder continuous on $\mathrm{dom}(\Psi)$ and $\mathrm{dom}(\Psi)$ is bounded provided the step-sizes θ_k are judiciously chosen. This is the same kind of convergence result shown in [11, Corollary 1] and in [4, Corollary 1].

Example 2. *The pair* $(\mathcal{D}, \mathsf{gap})$ *satisfies the* (q,r)*-growth property* (8) *for* q *and* $r = q/p$ *if* f *is* q*-uniformly smooth on* $\operatorname{dom}(\Psi)$ *and* Ψ *is* p*-uniformly convex on* $\operatorname{dom}(\Psi)$.
Detail. Suppose $q \in (1,2]$ and $p \geq 2$ are such that (22) holds for some $L > 0$ and (24) holds for some $\mu > 0$. Then for all $x \in \operatorname{dom}(\Psi)$ and $g := \nabla f(x)$ the tilted function $\Psi_g := \Psi + \langle g, \cdot \rangle$ is also p-uniformly convex on $\operatorname{dom}(\Psi_g) = \operatorname{dom}(\Psi)$ for the same p and μ. Consequently (25) implies that $s := \arg\min_y \{\Psi(y) + \langle g, y\rangle\} = \arg\min_y \Psi_g(y)$ satisfies

$$\mathsf{gap}(x,g) = \Psi(x) + \langle g, x\rangle - \Psi(s) - \langle g, s\rangle = \Psi_g(x) - \Psi_g(s) \geq \frac{\mu}{p}\|x - s\|^p.$$

Thus (23) implies that (8) holds for q and $r = q/p$ with $M = \frac{L}{q} \cdot \left(\frac{p}{\mu}\right)^{\frac{q}{p}}$. □

Observe that Ψ is 2-uniformly convex on $\operatorname{dom}(\Psi)$ if and only if Ψ is strongly convex on $\operatorname{dom}(\Psi)$. Thus Example 2 and Theorem 1 show that

$$\mathsf{gap}_k = \mathcal{O}(k^{\frac{2\nu}{\nu-1}})$$

when ∇f is ν-Hölder continuous on $\operatorname{dom}(\Psi)$ for some $\nu \in (0,1)$ and Ψ is strongly convex on $\operatorname{dom}(\Psi)$ provided the stepsizes θ_k are judiciously chosen. Furthermore, when $\nu = 1$, Example 2 and Theorem 1 also show that gap_k converges to zero *linearly*. These convergence rates match, modulo some slight log factor improvement, the convergence rates shown in [4, Corollary 1]. Example 2 and Theorem 1 imply the more general convergence rate

$$\mathsf{gap}_k = \mathcal{O}(k^{\frac{p\nu}{\nu+1-p}})$$

when ∇f is ν-Hölder continuous on $\operatorname{dom}(\Psi)$ for some $\nu \in (0,1]$ and Ψ is p-uniformly convex on $\operatorname{dom}(\Psi)$ for some $p > 2$ provided the stepsizes θ_k are judiciously chosen.

Example 3. *The pair* $(\mathcal{D}, \mathsf{gap})$ *satisfies the* (q,r)*-growth property* (8) *for* q *and* $r = q/p$ *if* f *is* q*-uniformly smooth on some* p*-uniformly convex and compact set* $C \subseteq \mathbb{R}^n$, $\Psi = \delta_C$, *and* ∇f *is bounded away from zero in* C.
Detail. Suppose $q \in (1,2]$ and $p \geq 2$ are such that (22) holds for some $L > 0$ and (26) holds for some $\mu > 0$. Then (27) implies that for $x \in \operatorname{dom}(\Psi), g := \nabla f(x)$ and $s = \arg\min_{x \in C} \langle g, x\rangle$ we have

$$\mathsf{gap}(x,g) = \langle g, x - s\rangle \geq \frac{\mu}{p}\|g\|^* \cdot \|x - s\|^p.$$

Thus (23) implies that (8) holds for q and $r = q/p$ with $M = \frac{L}{q} \cdot \left(\frac{p}{\ell \cdot \mu}\right)^{\frac{q}{p}}$ where

$$\ell = \inf_{x \in C} \|\nabla f(x)\| > 0.$$

□

Example 3 and Theorem 1 show that

$$\mathsf{gap}_k = \mathcal{O}(k^{\frac{p}{2-p}})$$

if f is 2-smooth on some p-uniformly convex and compact set $C \subseteq \mathbb{R}^n$ and $\Psi = \delta_C$ for some $p > 2$. Furthermore, when $p = 2$, Example 3 and Theorem 1 also show that gap_k converges to zero *linearly*. These convergence rates match the convergence rates shown in [7, Theorem 2.2].

4 Local growth property

We next describe an extension of the main developments in Section 2 can be extended to a *local* setting. Consider the same setup of Section 2. Suppose $q > 1$ and $r \in [0, 1]$. We shall say that the pair $(\mathcal{D}, \mathsf{gap})$ satisfies the (q, r)-*local growth property* if there exist constants $\delta > 0$ and $M > 0$ such that for all for all $x \in \mathrm{dom}(\Psi)$, $g := \nabla f(x)$, and $s := \partial\Psi^*(-g)$ that satisfy $\mathsf{gap}(x, g) < \delta$ we have

$$\mathcal{D}(x, s, \theta) \leq \frac{M\theta^q}{q} \cdot \mathsf{gap}(x, g)^r \text{ for all } \theta \in [0, 1]. \tag{28}$$

It is straightforward to see that local versions Theorem 1 and Proposition 2 also hold if we replace the growth property (8) with the local growth property (28) provided $\mathsf{gap}_0 = \mathsf{gap}(x_0, g_0) < \delta$. It is evident that the local growth property (28) is less stringent than the growth property (8). This is illustrated in the following example.

Example 4. *Suppose $x^\star$ is a minimizer of (1) and Ψ^* is differentiable in a neighborhood of $-\nabla f(x^\star)$. Suppose ∇f is Lipschitz continuous in a neighborhood of $x^\star$ and $\nabla\Psi^*$ is Lipschitz continuous in a neighborhood of $-\nabla f(x^\star)$. Then $(\mathcal{D}, \mathsf{gap})$ satisfies the (q, r)-local growth property for $q = 2, r = 1$. In particular, if gap_0 is sufficiently small then $\mathsf{gap}_k \to 0$ linearly.*

Detail. The above assumptions imply that f is 2-uniformly smooth and Ψ is 2-uniformly convex in some neighborhood of $x^\star$. They also imply that a point $x \in \mathrm{dom}(\Psi)$ belongs to that same neighborhood if $\mathsf{gap}(x, g)$ is sufficiently small. Proceeding as in Example 2 it follows that $(\mathcal{D}, \mathsf{gap})$ satisfies the (q, r)-local growth property for $q = 2$ and $r = 1$. Thus the local version of Theorem 1 implies that $\mathsf{gap}_k \to 0$ linearly provided gap_0 is sufficiently small. □

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] J. Borwein, G. Li, and L. Yao. Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets. *SIAM Journal on Optimization*, 24(1):498–527, 2014.

[3] R. Freund and P. Grigas. New analysis and results for the Frank–Wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.

[4] S. Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 173(1):431–464, 2019.

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

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

[7] T. Kerdreux, A. d'Aspremont, and S. Pokutta. Projection-free optimization on uniformly convex sets. In *International Conference on Artificial Intelligence and Statistics*, pages 19–27. PMLR, 2021.

[8] T. Kerdreux, L. Liu, S. Lacoste-Julien, and D. Scieur. Affine invariant analysis of Frank-Wolfe on strongly convex sets. In *International Conference on Machine Learning*, pages 5398–5408. PMLR, 2021.

[9] H. Lu. "Relative continuity" for non-Lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303, 2019.

[10] 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.

[11] Y. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1):311–330, 2018.

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

[13] Q. Van Nguyen. Variable quasi-Bregman monotone sequences. *Numerical Algorithms*, 73(4):1107–1130, 2016.

[14] Q. Van Nguyen. Forward-backward splitting with Bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.