Affine invariant convergence rates of the conditional gradient method

Javier F. Peña*

May 30, 2023

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 on the pair \((f, \Psi)\) in an affine invariant and norm-independent fashion.

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 on an open set containing \( \text{dom}(\Psi) \), 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 [7,12] corresponds to the case when \( \Psi \) 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. It is therefore natural to seek convergence results for

*Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu
the conditional gradient method that are affine invariant and norm-independent as any otherwise convergence result would typically be loose due to dependence on spurious factors.

We show that the conditional gradient method generates primal and dual iterates with a duality gap converging to zero provided a suitable \textit{growth property} holds (see (10)) and the algorithm makes a judicious choice of step-sizes (see (9) and (16)). The rate of convergence of the duality gap to zero ranges from sublinear to linear (see Theorem 1 and Theorem 2) depending on the degree of the growth property. The growth property and convergence results depend on the pair \((f, \Psi)\) in an affine invariant and norm-independent fashion. In fact, our main developments do not rely on any norms at all.

Our work is inspired by recent developments of Nesterov \cite{25}, Ghadimi \cite{10}, and Kerdreux et al \cite{14}. Under the assumption that the domain of \(\Psi\) is bounded, Nesterov \cite{25} established a \(O(1/k^\nu)\) convergence rate for the conditional gradient method when \(\nabla f\) is \(\nu\)-Hölder continuous. Nesterov \cite{25} also showed the stronger convergence rate \(O(1/k^{2\nu})\) under the additional assumption that \(\Psi\) is strongly convex. Ghadimi \cite{10} showed an even stronger linear rate of convergence when \(\nabla f\) is Lipschitz continuous and \(\Psi\) is strongly convex for a judicious choice of step-sizes. On the other hand, Kerdreux et al. \cite{14} considered the case when \(\nabla f\) is Lipschitz continuous and \(\Psi\) is the indicator function of a uniformly convex set. Kerdreux et al. \cite{14} showed that the convergence rate ranges from sublinear to linear depending on the degree of uniform convexity of the set. The results in \cite{10, 14, 25} are all norm-dependent since they rely on Hölder continuity and uniform convexity. The norm-dependence of the results in \cite{10, 14, 25} in turn implies that they are not affine invariant when the norms are not affine invariant, 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 Theorem 2) provided a \textit{growth property} (10) holds and the step-sizes are judiciously chosen. The growth property (10) can be seen as an extension of the \textit{curvature constant} proposed by Jaggi \cite{12}. It also has a flavor similar to that of the \textit{relative smoothness} and \textit{relative continuity} properties in the context of Bregman proximal methods as discussed in \cite{2, 22, 23, 28–30}. Our main convergence results (Theorem 1 and Theorem 2), obtained via the growth property (10), are stated in terms of the duality gap between the primal and dual iterates that the algorithm generates. We also develop parallel results (Theorem 4 and Theorem 5) on the suboptimality gap via a relaxation (40) of the growth property (10).

Aside from its natural compatibility with the properties of the conditional gradient method, the affine invariance and norm independence of our approach yields convergence results that are sharper and more general than previous approaches that do not have these two fundamental properties. More precisely, as we detail in Section 3, Section 4, and Section 5, our convergence results not only subsume some of the norm-dependent convergence results in \cite{8, 10, 14, 25, 31} but are also at least as sharp as these previous
norm-dependent results for any choice of norms. Put differently, the affine invariance and norm-independence of our main convergence results automatically circumvent any possible spurious overhead due to a particular choice of norms or affine scaling. It is also noteworthy that our main results (Theorem 1, Theorem 2, Theorem 3, and Theorem 4) rely only on straightforward stepsize procedures that are adaptive to the typically unknown parameters associated to the growth or weak growth properties. This stands in sharp contrast to some of the algorithmic schemes in [8,10,14,25,31] that rely on knowledge of specific parameters like Lipschitz continuity constants.

This article is related to a number of papers in the rich and increasingly growing literature on the conditional gradient method. Our primal-dual approach, with emphasis on bounding the duality gap for the conditional gradient iterates, is similar in spirit to the approaches used by Bach [1], Freund and Grigas [7], and Jaggi [12]. Like all of [1,7,12], our main results provide upper bounds on the duality gap between the (primal) conditional gradient iterates and some canonical dual iterates that the algorithm also generates. However, our central goal and main results are fundamentally different from those in [1,7,12]. Bach [1] focuses primarily on the equivalence between the conditional gradient and mirror descent algorithms. All of the convergence results in [1,7,12] are of order $O(1/k)$ or some minor variations. By contrast, our main focus is the spectrum of convergence rates of the conditional gradient method ranging from sublinear to linear. The latter goal is also pursued in the recent articles [10,13,25,31] albeit in a norm-dependent fashion. The articles [15,27] also establish a spectrum of convergence rates ranging from sublinear to linear but for more elaborate variants of the conditional gradient method that include various types of search directions, restart procedures, and/or are restricted to specific types of domains. As we detail in Section 3 and Section 4, our affine invariant and norm-independent results subsume, sharpen, and extend most of the norm-dependent results previously derived in [8,10,14,25,31]. Our developments, driven by affine-invariance, are in the same spirit as those of Kerdreux et al. [16]. However, the approach in [16] applies only to the case when $\Psi$ 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 $\Psi$ and relies on the growth property (10) that leads to convergence rates covering the entire range from sublinear to linear.

The remaining sections of the paper are organized as follows. Section 2 presents our main developments, namely the affine invariant and norm-independent growth property (10) 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 (9). Theorem 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 (16). Section 3 details some classes of problems that satisfy the growth property (10). As a byproduct, we recover, sharpen, and extend some of the norm-dependent convergence rates established in [10,14,25]. Section 4 presents developments parallel to those in Section 2 but for the suboptimality gap by relying on the weak version (10) of the growth
Section 5 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

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

(1)

where both $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ and $\Psi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ 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 an open set containing $\text{dom}(\Psi)$.

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

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

(2)

The format of problem (1) is the same as that considered in [10, 25]. The traditional format of the conditional gradient method, as discussed in [7, 12], corresponds to the special case when $\Psi$ is the indicator function $\Psi = \delta_C$ of some closed convex set $C \subseteq \mathbb{R}^n$. A fundamental difference between the conditional gradient method and other popular first-order methods for (1), like proximal gradient or more general Bregman proximal gradient methods, is the reliance on the linear oracle (2) as opposed to the reliance on a proximal mapping for the function $\Psi$. This fundamental difference has two profound implications. First, the conditional gradient method is preferable for a problem of the form (1) where the proximal mapping for $\Psi$ is unavailable or prohibitively expensive but the linear oracle (2) is available. This is the case in a large class of problems in data science such as those discussed in [1, 5–7, 9–12, 32]. In addition to the popular setting where $\Psi$ is the indicator function of a closed convex set, the articles [1, 5, 6, 9, 17, 20, 24, 32] also discuss problems of the form (1) where $\Psi$ is not an indicator function but a more involved regularization term. This more general version of the conditional gradient method is sometimes called generalized conditional gradient method. Second, the reliance on the linear oracle (2) as opposed to the reliance on proximal mappings endows the conditional gradient method with the affine invariance and norm-independence properties that play a central role in this article.

Our main developments (Theorem 1 and Theorem 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

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

(3)
Here \( f^* \) and \( \Psi^* \) denote the conjugates of \( f \) and \( \Psi \) 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 \( \Psi \) together with Assumptions A1 and A2 imply that \( \text{dom}(\Psi) \subseteq \text{dom}(f) \) and \( \text{dom}(f^*) \subseteq -\text{dom}(\Psi^*) \).

\begin{algorithm}
\caption{Conditional gradient algorithm}
1: \textbf{input:} \((f, \Psi), x_0 \in \text{dom}(\Psi)\)
2: \textbf{for} \( k = 0, 1, 2, \ldots \) \textbf{do}
3: \hspace{1em} \text{pick } s_k \in \arg \min_y \{ \langle \nabla f(x_k), y \rangle + \Psi(y) \} \text{ \ and } \theta_k \in [0, 1]
4: \hspace{1em} \text{let } x_{k+1} := (1 - \theta_k)x_k + \theta_k s_k
5: \textbf{end for}
\end{algorithm}

Starting from a current iterate \( x \), Algorithm 1 updates \( x \) to \( x_{k+1} := x + \theta(s - x) \) 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 \textit{Wolfe gap}:
\[
f(x) + \Psi(x) + f^*(g) + \Psi^*(-g) = \langle g, x - s \rangle + \Psi(x) - \Psi(s) \tag{4}
\]
or as
\[
f(s) + \Psi(s) + f^*(g) + \Psi^*(-g) = f(s) - f(x) - \langle g, s - x \rangle = D_f(s, x), \tag{5}
\]
where \( D_f \) denotes the \textit{Bregman distance} of \( f \), that is, the function
\[
D_f(y, x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle.
\]

In the special case when \( \Psi \) is the indicator function of a closed convex set \( C \subseteq \mathbb{R}^n \), the Wolfe gap (4) has the simpler and popular expression
\[
f(x) + \Psi(x) + f^*(g) + \Psi^*(-g) = \langle g, x - s \rangle.
\]

Our main developments are stated in terms of the functions \textit{gap} and \( D \) described next. Define the duality gap function \textit{gap} : \text{dom}(\Psi) \times \text{dom}(f^*) \rightarrow \mathbb{R} \) as follows
\[
\text{gap}(x, u) := f(x) + \Psi(x) + f^*(u) + \Psi^*(-u). \tag{6}
\]

Define \( D : \text{dom}(\Psi) \times \text{dom}(\Psi) \times [0, 1] \rightarrow \mathbb{R} \) as follows
\[
D(x, s, \theta) := D_f(x + \theta(s - x), x) + \Psi(x + \theta(s - x)) - (1 - \theta)\Psi(x) - \theta\Psi(s). \tag{7}
\]

The following proposition establishes a fundamental gap reduction identity connecting \textit{gap} and \( D \).
Proposition 1. Suppose $x \in \text{dom}(\Psi), g = \nabla f(x)$, and $s \in \partial \Psi^*(-g)$ Then for $\theta \in [0, 1]$ we have
\[
gap(x + \theta(s - x), g) = (1 - \theta)\gap(x, g) + D(x, s, \theta). \tag{8}\]

Proof. Adding (4) times $(1 - \theta)$ and (5) times $\theta$ we get
\[
(1 - \theta)(f + \Psi)(x) + \theta(f + \Psi)(s) + f^*(g) + \Psi^*(-g) = (1 - \theta)\gap(x, g) + \theta D_f(s, x)
\]
Thus
\[
\gap(x + \theta(s - x), g) = (f + \Psi)(x + \theta(s - x)) + f^*(g) + \Psi^*(-g)
\]
\[
= (1 - \theta)\gap(x, g) + \theta D_f(s, x) + (f + \Psi)(x + \theta(s - x)) - (1 - \theta)(f + \Psi)(x) - \theta(f + \Psi)(s).
\]
To finish, observe that
\[
\theta D_f(s, x) + (f + \Psi)(x + \theta(s - x)) - (1 - \theta)(f + \Psi)(x) - \theta(f + \Psi)(s)
\]
\[
= D_f(x + \theta(s - x), x) + \Psi(x + \theta(s - x)) - (1 - \theta)\Psi(x) - \theta\Psi(s)
\]
\[
= D(x, s, \theta).
\]

Theorem 1 below shows that Algorithm 1 generates primal and dual iterates whose duality gap converges to zero provided the growth property (10) 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 $\theta_k$ via the following line-search procedure
\[
\theta_k := \arg \min_{\theta \in [0, 1]} \{(1 - \theta)\gap(x_k, g_k) + D(x_k, s_k, \theta)\}. \tag{9}\]
As Theorem 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 (8) in Proposition 1 implies that (9) can also be equivalently stated as the exact line-search procedure
\[
\theta_k := \arg \min_{\theta \in [0, 1]} (f + \Psi)(x_k + \theta(s_k - x_k)).
\]
We prefer (9) because of its direct connection with the gap reduction identity (8).

Theorem 1 hinges on the following growth property connecting the functions $D$ and $\gap$. Suppose $q > 1$ and $r \in [0, 1]$. We shall say that the pair $(D, \gap)$ satisfies the $(q, r)$-growth property if there exists a finite constant $M > 0$ such that for all $x \in \text{dom}(\Psi), g := \nabla f(x)$, and $s \in \partial \Psi^*(-g)$
\[
D(x, s, \theta) \leq \frac{Mq^{\theta q}}{q} \cdot \gap(x, g)^r \text{ for all } \theta \in [0, 1]. \tag{10}\]
As we detail in Section 3 below, the \((q,r)\)-growth property (10) can be seen as a generalization of the curvature constant defined by Jaggi [12]. The growth property (10) 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 [2, 22, 23, 28–30].

The crux of Theorem 1 is to combine the gap reduction identity (8), the growth property (10), and the following technical lemma due to Borwein, Li, and Yao [4, Lemma 4.1].

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

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

Our two central results, namely Theorem 1 and Theorem 2, concern the sequence \( \text{gap}_k, k = 0, 1, \ldots \) of best duality gaps defined as follows. For \( k = 0, 1, \ldots \) let

\[
\text{gap}_k := \min_{i=0,1,\ldots,k} \text{gap}(x_k, g_i)
\]

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

Observe that \( \text{gap}_k = \text{gap}(x_k, g_k^{\text{best}}) \) where \( g_k^{\text{best}} \) is the best of the dual iterates \( g_i, i = 0, 1, \ldots, k \). A simple calculation shows that both \( \text{gap}_k \) and \( g_k^{\text{best}} \) can be updated as follows: \( \text{gap}_0 = \text{gap}(x_0, g_0) \) and for \( k = 0, 1, \ldots \)

\[
\text{gap}_{k+1} = \min\{\text{gap}(x_{k+1}, g_{k+1}), \text{gap}_k + (f + \Psi)(x_{k+1}) - (f + \Psi)(x_k)\},
\]

and

\[
g_k^{\text{best}} = \begin{cases} g_{k+1} & \text{if } \text{gap}(x_{k+1}, g_{k+1}) < \text{gap}_k + (f + \Psi)(x_{k+1}) - (f + \Psi)(x_k) \\ g_k^{\text{best}} & \text{otherwise.} \end{cases}
\]

**Theorem 1.** Suppose Algorithm 1 chooses \( \theta_k \in [0, 1] \) via (9) and \( q > 1 \) and \( r \in [0, 1] \) are such that \((D, \text{gap})\) satisfy the \((q,r)\)-growth property (10) for some finite \( M > 0 \). Then for \( k = 0, 1, \ldots \)

\[
\text{gap}_{k+1} \leq \text{gap}_k \left( 1 - \frac{q-1}{q} \cdot \min \left\{ 1, \left( \frac{\text{gap}_k^{1-r}}{M} \right)^{\frac{1}{q-1}} \right\} \right).
\]

When \( r = 1 \) we have linear convergence

\[
\text{gap}_k \leq \text{gap}_0 \left( 1 - \frac{q-1}{q} \cdot \min \left\{ 1, \frac{1}{M^{\frac{1}{q-1}}} \right\} \right)^k.
\]

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

\[
\text{gap}_k \leq \text{gap}_0 \left( 1 - \frac{q-1}{q} \right)^k, \quad k = 0, 1, 2, \ldots, k_0
\]
where \(k_0\) is the smallest \(k\) such that \(\text{gap}_k^{\frac{1}{1-r}} \leq M\). Then for \(k \geq k_0\) we have a sublinear convergence regime

\[
\text{gap}_k \leq \left( \text{gap}_{k_0}^{\frac{1}{1-r}} + \frac{1 - r}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}} \cdot (k - k_0) \right)^{\frac{q-1}{1}}. \tag{15}
\]

**Proof.** Identity (8) in Proposition 1 implies that

\[
\text{gap}(x_{k+1}, g_k) = (1 - \theta_k)\text{gap}(x_k, g_k) + D(x_k, s_k, \theta_k).
\]

Thus (11), the stepsize choice (9), and growth property (10) imply that

\[
\text{gap}_{k+1} \leq \text{gap}_k - \theta_k \cdot \text{gap}(x_k, g_k) + D(x_k, s_k, \theta_k) \\
= \text{gap}_k + \min_{\theta \in [0, 1]} \left\{ -\theta \cdot \text{gap}(x_k, g_k) + \frac{M \theta^q}{q} \cdot \text{gap}(x_k, g_k)^r \right\} \\
\leq \text{gap}_k + \min_{\theta \in [0, 1]} \left\{ -\theta \cdot \text{gap}(x_k, g_k) + \frac{M \theta^q}{q} \cdot \text{gap}(x_k, g_k)^r \right\} \\
\leq \text{gap}_k \left( 1 - \frac{q - 1}{q} \cdot \min \left\{ 1, \left( \frac{\text{gap}_k^{1-r}}{M} \right)^{\frac{1}{q-1}} \right\} \right).
\]

The last step above follows because the minimum in the second to last step is attained at

\[
\hat{\theta} = \min \left\{ 1, \left( \frac{\text{gap}(x_k, g_k)^{1-r}}{M} \right)^{\frac{1}{q-1}} \right\} \geq \min \left\{ 1, \left( \frac{\text{gap}_k^{1-r}}{M} \right)^{\frac{1}{q-1}} \right\}.
\]

Thus \(M \hat{\theta}^q \cdot \text{gap}(x_k, g_k)^r \leq \hat{\theta} \cdot \text{gap}(x_k, g_k)\) and consequently

\[
\text{gap}_k + \min_{\theta \in [0, 1]} \left\{ -\theta \cdot \text{gap}(x_k, g_k) + \frac{M \theta^q}{q} \cdot \text{gap}(x_k, g_k)^r \right\} \\
\leq \text{gap}_k - \hat{\theta} \cdot \text{gap}(x_k, g_k) + \frac{1}{q} \cdot \hat{\theta} \cdot \text{gap}(x_k, g_k) \\
\leq \text{gap}_k \left( 1 - \frac{q - 1}{q} \cdot \hat{\theta} \right) \\
\leq \text{gap}_k \left( 1 - \frac{q - 1}{q} \cdot \min \left\{ 1, \left( \frac{\text{gap}_k^{1-r}}{M} \right)^{\frac{1}{q-1}} \right\} \right).
\]

Therefore (12) is established. Inequality (12) automatically implies (13) when \(r = 1\) and also (14) for \(k \leq k_0\) when \(r \in [0, 1)\). When \(r \in [0, 1)\) and \(k \geq k_0\) inequality (15) follows from (12) and Lemma 1 applied to \(p := \frac{1-r}{q-1}\) and \(\beta_k := \text{gap}_k; \delta_k := \frac{q - 1}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}}\) 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 (11) yields a convergence rate for \(\text{gap}_k \rightarrow 0\) somewhere in the spectrum from
the sublinear rate $O(1/k^q)$ 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 \(10\) increases. Section 3 below elaborates on these observations and other interesting consequences of Theorem 1.

We next describe a more flexible alternative to \(9\) to choose the stepsize $\theta_k$. Suppose $c, \rho \in (0, 1)$. A simple backtracking procedure can easily choose $\theta_k \in [0, 1]$ such that

$$\theta_k^* := \max \{ \theta \in [0, 1] : (1 - \theta)\text{gap}(x_k, g_k) + D(x_k, s_k, \theta) \leq (1 - c \cdot \theta)\text{gap}(x_k, g_k) \}.$$ \(16\)

The inequality in \(16\) can be seen as an Armijo rule condition for the function $\varphi : [0, 1] \to \mathbb{R}$ defined as $\varphi(\theta) := (1 - \theta)\text{gap}(x_k, g_k) + D(x_k, s_k, \theta)$. Indeed, a simple calculation shows that $\varphi(0) = \text{gap}(x_k, g_k)$ and the right derivative of $\varphi$ at zero satisfies

$$\varphi'_+(0) \leq -\text{gap}(x_k, g_k) < 0.$$ 

Therefore the inequality in \(16\) can be rewritten in the usual Armijo rule format \[26\] Chapter 3:

$$\varphi(\theta) \leq \varphi(0) + \bar{c} \cdot \theta \cdot \varphi'_+(0).$$

for some $\bar{c} \leq c$. The Armijo rule condition \(16\) is evidently easier to satisfy, and therefore more flexible, than the exact line-search \(9\) which corresponds to choosing $\theta_k := \arg\min_{\theta \in [0, 1]} \varphi(\theta)$.

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

**Theorem 2.** Suppose $c, \rho \in (0, 1)$ are such that $c + \rho > 1$ and Algorithm 2 chooses $\theta_k \in [0, 1]$ so that $\rho \cdot \theta_k \leq \theta_k \leq \theta_k^*$ as in \(16\). Suppose $q > 1$ and $r \in [0, 1]$ are such that $(D, \text{gap})$ satisfy the $(q,r)$-growth property \(10\) for some finite $M > 0$. Then for $k = 0, 1, \ldots$

$$\text{gap}_{k+1} \leq \text{gap}_k \left(1 - (c + \rho - 1) \cdot \min \left\{ 1, \left( \frac{q(1-c)}{M} \cdot \text{gap}_k^{1-r} \right)^{\frac{1}{q-1}} \right\} \right).$$ \(17\)

When $r = 1$ we have linear convergence

$$\text{gap}_k \leq \text{gap}_0 \left(1 - (c + \rho - 1) \cdot \min \left\{ 1, \left( \frac{q(1-c)}{M} \right)^{\frac{1}{q-1}} \right\} \right)^k.$$ \(18\)

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

$$\text{gap}_k \leq \text{gap}_0 (1 - (c + \rho - 1))^k, \quad k = 0, 1, 2, \ldots, k_0$$ \(19\)

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

$$\text{gap}_k \leq \left( \text{gap}_{k_0} \cdot \left( \frac{(1-r)(c + \rho - 1)}{q - 1} \cdot \left( \frac{q(1-c)}{M} \right)^{\frac{1}{q-1}} \cdot (k - k_0) \right)^{\frac{q-1}{q-1}} \right)^{\frac{q-1}{q-1}}.$$ \(20\)
Proof. As noted above, this is a modification of the proof of Theorem 1. Identity (8) in Proposition 1 yields
\[ \text{gap}(x_{k+1}, g_k) = (1 - \theta_k)\text{gap}(x_k, g_k) + D(x_k, s_k, \theta_k). \]
Hence (11) implies that
\[ \text{gap}_{k+1} \leq \text{gap}_k - \theta_k \cdot \text{gap}(x_k, g_k) + D(x_k, s_k, \theta_k). \]
Since \( \rho \cdot \hat{\theta}_k \leq \theta_k \leq \hat{\theta}_k \) for \( \hat{\theta}_k \) as in (16), it follows that
\[
\begin{align*}
\text{gap}_{k+1} &\leq \text{gap}_k - \rho \cdot \hat{\theta}_k \cdot \text{gap}(x_k, g_k) + D(x_k, s_k, \theta_k) \\
&\leq \text{gap}_k - \rho \cdot \hat{\theta}_k \cdot \text{gap}(x_k, g_k) + (1 - c) \cdot \theta_k \cdot \text{gap}(x_k, g_k) \\
&\leq \text{gap}_k - (c + \rho - 1) \cdot \hat{\theta}_k \cdot \text{gap}(x_k, g_k) \\
&\leq \text{gap}_k(1 - (c + \rho - 1) \cdot \hat{\theta}_k).
\end{align*}
\] (21)

The growth property (10) and (11) imply that the minimizer \( \hat{\theta}_k \) in (16) satisfies
\[
\begin{align*}
\hat{\theta}_k &= \max \{ \theta \in [0, 1] : D(x_k, s_k, \theta) \leq (1 - c) \cdot \theta \cdot \text{gap}(x_k, g_k) \} \\
&\geq \max \left\{ \theta \in [0, 1] : \frac{M \theta^q}{q} \cdot \text{gap}(x_k, g_k)^r \leq (1 - c) \cdot \theta \cdot \text{gap}(x_k, g_k) \right\} \\
&= \min \left\{ 1, \left( \frac{q(1 - c)}{M} \cdot \text{gap}(x_k, g_k)^{1-r} \right)^{\frac{1}{q-1}} \right\} \\
&\geq \min \left\{ 1, \left( \frac{q(1 - c)}{M} \cdot \text{gap}_k^{1-r} \right)^{\frac{1}{q-1}} \right\}. \quad (22)
\end{align*}
\]

Inequality (17) follows by putting together (21) and (22). Inequality (17) automatically implies (18) when \( r = 1 \) and also (19) for \( k \leq k_0 \) when \( r \in [0, 1) \). When \( r \in [0, 1) \) and \( k \geq k_0 \) inequality (20) follows from (17) and Lemma 1 applied to \( p := \frac{1-r}{q-1} \) and \( \beta_k := \text{gap}_k, \delta_k := (c + \rho - 1) \left( \frac{q(1-c)}{M} \right)^{\frac{1}{q-1}} \) for \( k = k_0, k_0 + 1, \ldots \).

We note that expression (17) in Theorem 2 is identical to expression (12) in Theorem 1 for the ideal choice \( \rho = 1 \) and \( c = (q-1)/q \). Likewise for (18), (20) and (13), (15).

To conclude this section, we next discuss the norm-independence and affine invariance of our developments. The growth property (10) as well as Theorem 1 and Theorem 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 \( x := A \bar{x} + b \) where \( A : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a linear bijection. Let \( \tilde{f}, \tilde{\Psi} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{ \infty \} \) be defined as the corresponding reparametrizations of \( f \) and \( \Psi \), namely
\[ \tilde{f}(\bar{x}) := f(A \bar{x} + b) \quad \text{and} \quad \tilde{\Psi}(\bar{x}) := \Psi(A \bar{x} + b). \]
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

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

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

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

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 \( \partial \tilde{\Psi}^*(-g) = A \partial \tilde{\Psi}^*(-\tilde{g}) + b \). Thus we can assume that the oracle in Assumption A2 yield respectively \( s \in \partial \tilde{\Psi}^*(-g) \) 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

\[
\text{gap}(x, g) = \text{gap}(\tilde{x}, \tilde{g})
\]

and

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

It thus follows that the growth property (10), stepsizes selection procedures (9) and (16), as well as Theorem 1 and Theorem 2 are all affine invariant.

### 3 Growth property in some special cases

The subsections below detail some straightforward consequences of Theorem 1 when the growth property holds for some special values of \((q, r)\). These special cases recover and sharpen some norm-dependent convergence results previously developed in the state-of-the-art literature on the conditional gradient method [10, 14, 25]. To make the latter connection more precise and explicit, Proposition 2, Proposition 3, and Proposition 4 below show that the uniform smoothness and uniform convexity assumptions previously made in each of [10, 14, 25] imply the \((q, r)\) growth property (10) for various values of \((q, r)\).

We next recall the concepts of uniform smoothness and uniform convexity. A detailed discussion on these concepts is presented in [13]. Both uniform smoothness and uniform convexity are stated and formalized in terms of some norm. Thus throughout this section we assume that \( \mathbb{R}^n \) is endowed with a norm \( \| \cdot \| \) and let \( \| \cdot \|^{*} \) denote its dual norm.

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

\[
f(x + \theta(y - x)) \geq (1 - \theta)f(x) + \theta f(y) - \frac{L}{q} \theta(1 - \theta)\|y - x\|^q \quad \text{for all} \quad \theta \in [0, 1]. \tag{23}
\]

It is customary to say that \( f \) is \( L \)-smooth on \( C \) when (23) holds with \( q = 2 \). In particular, **uniform smoothness** is a relaxation of **smoothness**. It is easy to see that (23) holds for \( q = \nu + 1 \) when \( \nabla f \) is \( \nu \)-Hölder continuous on \( C \) with Hölder constant \( L \). In particular, \( f \) is \( L \)-smooth on \( C \) when \( \nabla f \) is Lipschitz continuous on \( C \) with Lipschitz constant \( L \).
As detailed in [13], when \( f \) is differentiable, property (23) implies that for all \( x, y \in C \)
\[
D_f(y, x) \leq \frac{L}{q} \|y - x\|^q.
\]
(24)

A function \( \Psi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{ \infty \} \) is \textit{uniformly convex} on \( C \subseteq \text{dom}(\Psi) \) with respect to \( \| \cdot \| \) if there exist constants \( p \geq 2 \) and \( \mu > 0 \) such that for all \( x, y \in C \) and \( \theta \in [0, 1] \)
\[
\Psi(x + \theta(y - x)) \leq (1 - \theta)\Psi(x) + \theta\Psi(y) - \frac{\mu}{p}\theta(1 - \theta)\|y - x\|^p \text{ for all } \theta \in [0, 1].
\]
(25)

It is customary to say that \( \Psi \) is strongly convex on \( C \) when (25) holds for \( p = 2 \). In particular, \textit{uniform convexity} is a relaxation of \textit{strong convexity}.

Again, as detailed in [13], property (25) implies that for all \( x, y \in C \) and \( g \in \partial \Psi(y) \)
\[
\Psi(x) \geq \Psi(y) + \langle g, s - y \rangle + \frac{\mu}{p}\|x - y\|^p.
\]
(26)

In particular, property (25) implies that if \( C = \text{dom}(\Psi) \) and \( s = \arg \min_y \Psi(y) \) then for all \( x \in \text{dom}(\Psi) \)
\[
\Psi(x) - \Psi(s) \geq \frac{\mu}{p}\|x - s\|^p.
\]
(26)

A closed convex set \( C \subseteq \mathbb{R}^n \) is \textit{uniformly convex} if there exist constants \( p \geq 2 \) and \( \mu > 0 \) such that for all \( x, y \in C, \theta \in [0, 1], \) and \( z \in \mathbb{R}^n \) with \( \|z\| \leq 1 \)
\[
x + \theta(y - x) + \frac{\mu}{p}\theta(1 - \theta)\|y - x\|^pz \in C.
\]
(27)

It is customary to say that \( C \) is strongly convex when (27) holds for \( p = 2 \).

Similar to the above implications (23) \( \Rightarrow \) (24) and (25) \( \Rightarrow \) (26), property (27) implies that if \( g \in \mathbb{R}^n \) and \( s = \arg \min_{y \in C} \langle g, y \rangle \) then for all \( x \in C \)
\[
\langle g, x - s \rangle \geq \frac{\mu}{p}\|g\|_{\ast} \cdot \|x - s\|^p.
\]
(28)

### 3.1 The case \( q > 1, r = 0 \): sublinear convergence

Suppose the growth property (10) holds for \( q > 1 \) and \( r = 0 \), that is, for \( x \in \text{dom}(\Psi), s \in \partial \Psi^*(-\nabla f(x)) \)
\[
D(x, s, \theta) \leq \frac{M\theta^q}{q} \text{ for all } \theta \in [0, 1].
\]
(29)

In this case \( \text{gap}_1 \leq M/q \leq M \) and thus Theorem \( \Pi \) implies that
\[
\text{gap}_k \leq \left( \frac{\text{gap}_1}{q} + \frac{1}{q} \cdot \frac{k - 1}{M \frac{1}{q - 1}} \right)^{-(q - 1)}
\leq \left( \frac{q}{M \frac{1}{q - 1}} + \frac{1}{q} \cdot \frac{k - 1}{M \frac{1}{q - 1}} \right)^{-(q - 1)}
= M \left( \frac{q}{k - 1 + q \frac{1}{q - 1}} \right)^{q - 1}.
\]
(30)
The case \((q, r) = (2, 0)\) is particularly interesting. In this case we get

\[
\text{gap}_k \leq \frac{2M}{k+3}. \tag{31}
\]

This bound applies in the especially important and popular case when \(\Psi\) is the indicator function \(\delta_C\) of some compact convex set \(C \subseteq \mathbb{R}^n\) as discussed in \([7, 12]\). In this case for \(x \in \text{dom}(\Psi), s \in \partial \Psi^*(-\nabla f(x))\) we have

\[
\mathcal{D}(x, s, \theta) = D_f(x + \theta(s - x), x)
\]

and thus \((29)\) holds with \(q = 2\) and \(M\) equal to the following curvature constant defined by Jaggi \([12]\), provided that it is finite:

\[
C_f := \sup_{x, s \in C, \theta \in [0, 1]} \frac{2D_f(x + \theta(s - x), x)}{\theta^2}. \tag{32}
\]

Therefore \((31)\) recovers the iconic convergence rate of the conditional gradient method established in \([12]\). However, Theorem \(1\) reveals that the convergence could be faster as \((29)\) may hold for \(M \leq C_f\).

Proposition 2. The \((q, 0)\)-growth property \((29)\) holds if \(f\) is \(q\)-uniformly smooth and \(\text{dom}(\Psi)\) is bounded. More precisely, the \((q, 0)\)-growth property \((29)\) holds for \(M = L \cdot \text{diam}(\text{dom}(\Psi))^q\) if there is a norm \(\| \cdot \|\) in \(\mathbb{R}^n\) such that \((23)\) holds for \(x, y \in \text{dom}(\Psi)\).

**Proof.** Suppose \((23)\) holds for some constants \(q > 1\) and \(L > 0\). Then \((7)\) and \((24)\) imply that for all \(x, s \in \text{dom}(\Psi)\) and \(\theta \in [0, 1]\)

\[
\mathcal{D}(x, s, \theta) \leq D_f(x + \theta(s - x), x) \leq \frac{L}{q} \|\theta(s - x)\|^q = \frac{L\|s - x\|^q}{q} \cdot \theta^q.
\]

Thus \((29)\) holds for \(M = L \cdot \text{diam}(\text{dom}(\Psi))^q\). \(\square\)

Proposition 2 and \((30)\) readily yield the following norm-dependent convergence result for a class of problems previously considered in \([10, 25]\). Suppose \(\text{dom}(\Psi)\) is bounded and \(\nabla f\) is \(\nu\)-Hölder continuous on \(\text{dom}(\Psi)\) for some \(\nu > 0\). In this case \(f\) is \((\nu + 1)\)-uniformly smooth on \(\text{dom}(\Psi)\). Thus Proposition 2 implies that the \((q, 0)\)-growth property \((29)\) holds for \(q = \nu + 1\) and \(M = L \cdot \text{diam}(\text{dom}(\Psi))^{1+\nu}\) where \(L\) is the Hölder continuity constant of \(\nabla f\). Therefore \((30)\) implies that

\[
\text{gap}_k \leq L \cdot \text{diam}(\text{dom}(\Psi))^{1+\nu} \left(\frac{\nu + 1}{k - 1 + (\nu + 1)^{\frac{1}{\nu}}}ight)^{\nu} \\
\leq L \cdot \text{diam}(\text{dom}(\Psi))^{1+\nu} \left(\frac{\nu + 1}{k}ight)^{\nu}. \tag{33}
\]
Therefore we recover the $O(1/k^{\nu})$ convergence results shown in [25, Corollary 1] and in [10, Corollary 1] with an improvement on the latter as [33] does not include the additional log factor in [10, Corollary 1 and Equation (2.30)]. The constant in front of $1/k^{\nu}$ in (33) is simpler and sharper than the constants in [25, Corollary 1] and in [10, Equation (2.30)]. Furthermore, Theorem 1 reveals that the convergence is generally faster as Proposition 2 implies that (29) and (30) will typically hold for $M \leq L \cdot \text{diam}(\text{dom}(\Psi))^{1+\nu}$.

### 3.2 The case $(q, r) = (2, 1)$: linear convergence

Suppose the growth property (10) holds for $(q, r) = (2, 1)$, that is, for $x \in \text{dom}(\Psi), s \in \partial\Psi^*(-\nabla f(x))$

$$D(x, s, \theta) \leq \frac{M \theta^2}{2} \cdot \text{gap}(x, g) \text{ for all } \theta \in [0, 1].$$

(34)

In this case Theorem 1 readily implies that

$$\text{gap}_k \leq \text{gap}_0 \cdot \left(1 - \frac{1}{2} \cdot \min \left\{1, \frac{1}{M}\right\}\right)^k.$$  

(35)

Proposition 3 below describes some important classes of problems that satisfies (34). We omit the proof of Proposition 3 as it is a special case of the more general Proposition 4 in Section 3.3 below.

**Proposition 3.** Suppose $f$ is 2-uniformly smooth on $\text{dom}(\Psi)$. Then the $(2, 1)$-growth property (34) holds if $\Psi$ is 2-uniformly convex on $\text{dom}(\Psi)$, or if $\Psi = \delta_C$ for a 2-uniformly convex and closed set $C \subseteq \mathbb{R}^n$ and $\nabla f$ is bounded away from zero in $C$. More precisely:

(a) The $(2, 1)$-growth property (34) holds for $M = 2L/\mu$ if there is a norm $\| \cdot \|$ in $\mathbb{R}^n$ such that both (23) and (25) hold for $p = q = 2$ and $x, y \in \text{dom}(\Psi)$.

(b) The $(2, 1)$-growth property (34) holds for $M = 2L/(\ell \mu)$ if $\Psi = \delta_C$ for a closed convex set $C \subseteq \mathbb{R}^n$, there is a norm $\| \cdot \|$ in $\mathbb{R}^n$ such that both (23) and (27) hold for $p = q = 2$ and $x, y \in C$, and

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

Proposition 3 and (35) readily yield the following norm-dependent linear convergence results for two classes of problems previously considered in [10,16,19].

First, if $\nabla f$ is Lipschitz continuous on $\text{dom}(\Psi)$ and $\Psi$ is strongly convex then Proposition 3(a) implies that the $(2, 1)$-growth property (34) holds with $M = L/\mu$ where $L$ and $\mu$ are respectively the Lipschitz continuity constant of $\nabla f$ and the strong convexity constant of $\Psi$. Therefore (35) implies that

$$\text{gap}_k \leq \text{gap}_0 \cdot \left(1 - \frac{1}{2} \cdot \min \left\{1, \frac{\mu}{2L}\right\}\right)^k.$$
Hence we recover the linear convergence result shown in [10, Corollary 1 and Equation (2.32)] but with a simpler and sharper expression. Furthermore, the convergence is generally faster as \( M \leq 2L/\mu \).

Second, if \( \Psi = \delta_C \), \( \nabla f \) is Lipschitz continuous and bounded away from zero on \( C \), and \( C \subseteq \mathbb{R}^n \) is a closed and strongly set then Proposition [31(b)] implies that the (2,1)-growth property \( (34) \) holds with \( M = 2L/(\ell \mu) \) where \( L, \mu, \) and \( \ell \) are respectively the Lipschitz continuity constant of \( \nabla f \), the strong convexity constant of \( C \) on \( C \). Therefore \( (35) \) implies that

\[
\text{gap}_k \leq \text{gap}_0 \cdot \left( 1 - \frac{1}{2} \cdot \min \left\{ 1, \frac{\ell \mu}{2L} \right\} \right)^k.
\]

Hence we recover the linear convergence result shown in [19, Theorem 6.1(5)] and in [16, Theorem 2.2] with a slightly improved convergence rate. Theorem 1 shows that the convergence rate is generally better as \( M \leq 2L/(\ell \mu) \).

### 3.3 Other values \((q, r)\): sublinear to linear convergence

Suppose the growth property \( (10) \) holds for some \( q > 1 \), \( r \in (0, 1) \), and \( M > 0 \). We next discuss the spectrum of convergence rates from sublinear to linear that Theorem 1 yields depending on the values of \((q, r)\). To simplify notation and exposition, we shall assume that \( \text{gap}_0 \leq M \) throughout this subsection so that we skip the initial linear convergence regime \( (14) \) in Theorem 1. This assumption can be relaxed either at the expense of some notational overhead to deal with the initial linear convergence regime \( (14) \), or at the expense of artificially increasing the constant \( M \).

Suppose \( \text{gap}_0 \leq M \). Then for \( r \in [0, 1) \) Theorem 1 implies that

\[
\text{gap}_k \leq \left( \frac{1}{M^{\frac{1}{q-1}}} + \frac{1-r}{q} \cdot \frac{k}{M^{\frac{1}{q-1}}} \right)^{\frac{q-1}{q-1}} \cdot M \left( 1 + \frac{q-1}{q \cdot M^{\frac{1}{q-1}}} \cdot k \cdot \frac{1-r}{q-1} \right)^{\frac{q-1}{q-1}}. \tag{36}
\]

For \( q > 1 \) fixed, the right-most expression in (36) decreases monotonically with \( r \) from the sublinear rate \( M \left( \frac{q}{k+q} \right)^{q-1} \) for \( r = 0 \) to the following asymptotic linear rate for \( r \uparrow 1 \)

\[
Me^{-\rho k} = M \left( e^{-\rho} \right)^k, \tag{37}
\]

where

\[
\rho := \frac{q-1}{q \cdot M^{\frac{1}{q-1}}} \in (0, 1).
\]

Since \( e^{-\rho} \approx 1 - \rho \) for small \( \rho > 0 \), it follows that the asymptotic linear rate (37) is approximately the same linear rate given by Theorem 1 when \( r = 1 \), namely

\[
\text{gap}_k \leq \text{gap}_0 \left( 1 - \frac{q-1}{q \cdot M^{\frac{1}{q-1}}} \right)^k.
\]
The following general version of Proposition 3 describes a broad class of problems that satisfies the \((q,r)\)-growth property for various values of \((q,r)\) depending on the degree of uniform smoothness and uniform convexity of \(f\) and \(\Psi\).

**Proposition 4.** Suppose \(f\) is \(q\)-uniformly smooth on \(\text{dom}(\Psi)\). Then the \((q,r)\)-growth property \((10)\) holds for \(r = q/p\) if \(\Psi\) is \(p\)-uniformly convex on \(\text{dom}(\Psi)\), or if \(\Psi = \delta_C\) for a \(p\)-uniformly convex and closed set \(C \subseteq \mathbb{R}^n\) and \(\nabla f\) is bounded away from zero in \(C\).

More precisely:

(a) The \((q,r)\)-growth property \((10)\) holds for \(r = q/p\) and \(M = L \cdot (p/\mu)^{q/p}\) if there is a norm \(\| \cdot \|\) in \(\mathbb{R}^n\) such that both \((23)\) and \((25)\) hold for \(x, y \in \text{dom}(\Psi)\).

(b) The \((q,r)\)-growth property \((10)\) holds for \(r = q/p\) and \(M = L \cdot (p/\ell\mu)^{q/p}\) if \(\Psi = \delta_C\) for a closed convex set \(C \subseteq \mathbb{R}^n\), and there is a norm \(\| \cdot \|\) in \(\mathbb{R}^n\) such that both \((23)\) and \((27)\) hold for \(x, y \in C\), and

\[
\ell := \inf_{x \in C} \| \nabla f(x) \|^* > 0.
\]

**Proof.** (a) Suppose \((23)\) holds for some constants \(q > 1\) and \(L > 0\), and also \((24)\) holds for some constants \(p > 1\) and \(\mu > 0\). Then for all \(x \in \text{dom}(\Psi)\) and \(g := \nabla f(x)\) the tilted function \(\Psi_g := \Psi + \langle g, \cdot \rangle\) is also \(p\)-uniformly convex on \(\text{dom}(\Psi_g) = \text{dom}(\Psi)\) for the same \(p\) and \(\mu\). Consequently \((26)\) implies that \(s := \arg \min_y \{ \Psi(y) + \langle g, y \rangle \} = \arg \min_y \Psi_g(y)\) satisfies

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

On the other hand, \((7)\) and \((24)\) imply that

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

It thus follows that \((10)\) holds for \(q\) and \(r = q/p\) with \(M = L \cdot (p/\mu)^{q/p}\).

(b) Suppose \((23)\) holds for some constants \(q > 1\) and \(L > 0\), \((24)\) holds for some constants \(p \geq 2\) and \(\mu > 0\), and \(\ell := \inf_{x \in C} \| \nabla f(x) \|^* > 0\). Then \((28)\) implies that for \(x \in \text{dom}(\Psi)\), \(g := \nabla f(x)\) and \(s = \arg \min_{x \in C} \langle g, x \rangle\) we have

\[
\text{gap}(x, g) = \langle g, x - s \rangle \geq \frac{\mu}{p} \| g \|^* \cdot \| x - s \|^p \geq \frac{\ell \mu}{p} \cdot \| x - s \|^p.
\]

Thus proceeding as in part (a), \((24)\) implies that \((10)\) holds for \(q\) and \(r = q/p\) with \(M = L \cdot (p/(\ell \mu))^{q/p}\). \(\square\)
Proposition 4 and Theorem 1 yield the following norm-dependent convergence results for two classes of problems that extend the two classes discussed in Section 3.2.

First, suppose that \( \nabla f \) is \( \nu \)-Hölder continuous on \( \text{dom}(\Psi) \) for some \( \nu \in (0, 1] \) and \( \Psi \) is \( p \)-uniformly convex on \( \text{dom}(\Psi) \) for some \( p \geq 2 \). Then Proposition 4 implies that the \((q, r)\)-growth property (10) holds for \( q = \nu + 1 \) and \( r = (\nu + 1)/p \) with \( M = L \cdot (p/\mu)^{\frac{2}{\nu+1}} \).

The case \( r = 1 \), that is, \( \nu = 1 \) and \( p = 2 \) was already discussed in Section 3.2. Hence we focus on the case \( r < 1 \), that is, \( p > \nu + 1 \). Under the mild assumption that \( \text{gap} \leq M \), inequality (36) implies that

\[
\text{gap}_k \leq \left( \frac{1}{M^{\frac{p-\nu-1}{p}} + \frac{p-\nu-1}{p(\nu+1)} \cdot \frac{k}{M^{\frac{1}{p}}}} \right)^{\frac{p}{p-\nu-1}} \leq \left( \frac{p(\nu+1) \cdot M^{\frac{1}{p}}}{p-\nu-1} \cdot \frac{1}{k} \right)^{\frac{p}{p-\nu-1}}.
\]  

(39)

The above \( \mathcal{O}(1/k^{\frac{p}{p-\nu-1}}) \) convergence rate is new for \( p > 2 \) and extends some of the results considered in [10, 25] that only consider the cases when \( p = 2 \) or when \( \text{dom}(\Psi) \) is bounded.

Second, if \( \Psi = \delta_C \), \( \nabla f \) is \( \nu \)-Hölder continuous and bounded away from zero on \( C \), and \( C \subseteq \mathbb{R}^n \) is a closed and \( p \)-uniformly set then Proposition 4(b) implies that the \((q, r)\)-growth property (34) holds for \( q = \nu + 1 \), \( r = \frac{q}{p} \) with \( M = L \cdot (p/(\ell \mu))^{\frac{1}{2}} \) where \( L, \mu, \) and \( \ell \) are respectively the Lipschitz continuity constant of \( \nabla f \), the strong convexity constant of \( C \) and the lower bound on \( \| \nabla f(x) \|^* \) on \( C \). Again the case \( r = 1 \), that is, \( \nu = 1 \) and \( p = 2 \) was already discussed in Section 3.2 and thus we focus on the case \( r < 1 \). Under the mild assumption that \( \text{gap} \leq M \), inequality (36) again implies (39) albeit for a different constant \( M \). Hence we extend the sublinear convergence result shown in [14, Theorem 2.2] that considers only the special case \( \nu = 1 \) and \( p > 2 \). For these values of \( \nu \) and \( p \) we get

\[
\text{gap}_k \leq \left( \frac{2Mp}{p-2} \cdot \frac{1}{k} \right)^{\frac{p}{p-2}}
\]

and thus recover the \( \mathcal{O}(1/k^{\frac{p}{p-2}}) \) sublinear convergence result in [14, Theorem 2.2] albeit with a slightly simpler and sharper constant.

4 Weak growth property and suboptimality gap

Consider the setting of Section 2. Define the suboptimality gap function \( \text{subopt} : \text{dom}(\Psi) \to \mathbb{R} \) as follows

\[
\text{subopt}(x) := f(x) + \Psi(x) - \min_y \{ f(y) + \Psi(y) \}.
\]

The weak duality property between (1) and its dual (3) implies that for all \((x, u) \in \text{dom}(\Psi) \times \text{dom}(f^*)\)

\[
\text{gap}(x, u) \geq \text{subopt}(x).
\]

Therefore any upper bound on the duality gap automatically holds for the suboptimality gap as well. This section takes this observation further: Theorem 3 and Theorem 4
provide analogous statements to Theorem 1 and Theorem 2 for the suboptimality gap by relying on the following relaxed version of (10). Suppose $q > 1$ and $r \in [0, 1]$. We shall say that the triple $(\mathcal{D}, \text{gap}, \text{subopt})$ satisfies the $(q, r)$ weak growth property if there exists a finite constant $M > 0$ such that for all $x \in \text{dom}(\Psi)$, $g := \nabla f(x)$, and $s \in \partial \Psi^*(-g)$

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

(40)

Since $\text{gap}(x, g) \geq \text{subopt}(x) \geq 0$, it is immediate that the weak growth property (40) holds whenever the growth property (10) holds.

The next two results concern the sequence of suboptimality gaps

$$
\text{subopt}_k := \text{subopt}(x_k), \ k = 0, 1, \ldots
$$

for the iterates generated by Algorithm 1. We omit the proofs of Theorem 3 and Theorem 4 below as they are straightforward extensions of the proofs of Theorem 1 and Theorem 2.

**Theorem 3.** Suppose Algorithm 1 chooses $\theta_k \in [0, 1]$ via (9) and $q > 1$ and $r \in [0, 1]$ are such that $(\mathcal{D}, \text{gap}, \text{subopt})$ satisfy the $(q, r)$ weak growth property (40) for some finite $M > 0$. Then for $k = 0, 1, \ldots$

$$
\text{subopt}_{k+1} \leq \text{subopt}_k \left(1 - \frac{q-1}{q} \cdot \min \left\{1, \left(\frac{\text{subopt}_k^{1-r}}{M} \right)^{\frac{1}{q-r}} \right\}\right).
$$

When $r = 1$ we have linear convergence

$$
\text{subopt}_k \leq \text{subopt}_0 \left(1 - \frac{q-1}{q} \cdot \min \left\{1, \frac{1}{M^{\frac{1}{q-1}}} \right\}\right)^k.
$$

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

$$
\text{subopt}_k \leq \text{subopt}_0 \left(1 - \frac{q-1}{q}\right)^k, \ k = 0, 1, 2, \ldots, k_0
$$

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

$$
\text{subopt}_k \leq \left(\text{subopt}_{k_0}^{\frac{r-1}{q-1}} + \frac{1-r}{q} \cdot \frac{1}{M^{\frac{1}{q-1}}} \cdot (k - k_0)\right)^{\frac{q-1}{r-1}}.
$$

**Theorem 4.** Suppose $c, \rho \in (0, 1)$ are such that $c + \rho > 1$ and Algorithm 1 chooses $\theta_k \in [0, 1]$ so that $\rho \cdot \delta_k \leq \theta_k \leq \delta_k$ for $\delta_k$ as in (10). Suppose $q > 1$ and $r \in [0, 1]$ are such that $(\mathcal{D}, \text{gap}, \text{subopt})$ satisfy the $(q, r)$ weak growth property (40) for some finite $M > 0$. Then for $k = 0, 1, \ldots$

$$
\text{subopt}_{k+1} \leq \text{subopt}_k \left(1 - (c + \rho - 1) \cdot \min \left\{1, \left(\frac{q(1-c)}{M} \cdot \text{subopt}_k^{1-r}\right)^{\frac{1}{q-1}} \right\}\right).
$$
When $r = 1$ we have linear convergence

$$\text{subopt}_k \leq \text{subopt}_0 \left(1 - (c + \rho - 1) \cdot \min \left\{1, \left(\frac{q(1 - c)}{M}\right)^{\frac{q-1}{q}}\right\}\right)^k.$$ 

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

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

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

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

Proposition 5 below provides a nice complement to Proposition 4(b). Proposition 5 describes a class of problems that satisfies the weak growth property similar to that in Proposition 4(b). However, instead of assuming that $\nabla f$ is bounded away from zero in $C$, Proposition 5 relies on the following type of Hölderian error bound as in [14,31]. Suppose $C \subseteq \mathbb{R}^n$ is a closed convex set, $f^* := \min_{x \in C} f(x)$ is finite, $X^* := \{x \in C : f(x) = f^*\}$ is nonempty. Suppose $\mathbb{R}^n$ is endowed with a norm $\| \cdot \|$ and $\gamma \in [0, 1]$. We shall say that $f$ satisfies the $\gamma$-Hölderian error bound on $C$ if there exists a finite constant $K > 0$ such that for all $x \in C$

$$\min_{y \in X^*} \|x - y\| \leq K \cdot (f(x) - f^*)^{\gamma}. \quad (41)$$

The Hölderian error bound (41) is known to hold generically when $C$ is a compact set and $f$ is subanalytic, as established by Lojasiewicz [21], and further developed by Kurdyka [18] and Bolte et al. [3].

**Proposition 5.** The $(q, r)$ weak growth property (40) holds for $r = \gamma q/p$ if $\Psi = \delta_C$ for some closed and $p$-uniformly convex set $C \subseteq \mathbb{R}^n$, $f$ is $q$-uniformly smooth on $C$, and the $\gamma$-Hölderian error bound (41) holds for $\gamma \in [0, 1]$.

**Proof.** Suppose (23) holds for some constants $q \in (1, 2]$ and $L > 0$, (27) holds for some constants $p \geq 2$ and $\mu > 0$, and (41) holds for some constants $\gamma \in [0, 1]$ and $K > 0$. Then (28) implies that for all $x \in C$ and $g := \nabla f(x)$ we have

$$\text{gap}(x, g) = \langle g, x - s \rangle \geq \frac{\mu}{p} \|g\|^p \cdot \|x - s\|^p. \quad (42)$$

Next, the convexity of $f$, Cauchy-Schwarz inequality, and (41) imply that for $x^* := \arg \min_{y \in X^*} \|x - y\|

$$\text{subopt}(x) = f(x) - f(x^*) \leq \langle g, x - x^* \rangle \leq \|g\|^p \cdot \|x - x^*\| \leq K \cdot \|g\|^p \cdot \text{subopt}(x)^\gamma. \quad (43)$$
From (42) and (43) it follows that
\[
\|x - s\|^p \cdot \text{subopt}(x)^{1-\gamma} \leq \frac{pK}{\mu} \cdot \text{gap}(x, g).
\]

The latter inequality, (24), and \(\text{subopt}(x) \leq \text{gap}(x, g)\) imply that
\[
D(x, s, \theta) \cdot \text{subopt}(x)^{1-\gamma} \leq \frac{L}{q} \cdot \|s - x\|^q \cdot \text{subopt}(x)^{\frac{(1-\gamma)q}{p}} \cdot \text{gap}(x, g)^{1-\gamma}.
\]

Therefore (40) holds for \(r = \frac{\gamma q}{p}\) and \(M = L \cdot \left(\frac{pK}{\mu}\right)^{q/p}\).

Proposition 5 and Theorem 3 yield the following norm-dependent convergence results for a class of problems previously considered in [16, 31].

Suppose \(\Psi = \delta_C\) for \(C \subseteq \mathbb{R}^n\) closed and \(p\)-uniformly convex, \(\nabla f\) is \(\nu\)-Hölder continuous on \(C\), and \(f\) satisfies the \(\gamma\)-Hölderian error bound on \(C\). Then Proposition 5 implies that the \((q, r)\) weak growth property holds for \(q = \nu + 1, r = \frac{\gamma q}{p}\) with \(M = L \cdot \left(\frac{pK}{\mu}\right)^{q/p}\) where \(L, \mu,\) and \(K\) are respectively the Hölder continuity constant of \(\nabla f\), the \(p\)-convexity constant of \(C\), and the \(\gamma\)-Hölderian error bound constant of \(f\) on \(C\).

When \(\nu = 1, p = 2,\) and \(\gamma = 1\) we have \(q = 2, r = 1,\) and \(M = 2LK\). Thus Theorem 3 yields the linear convergence rate
\[
\text{subopt}_k \leq \text{subopt}_0 \left(1 - \frac{1}{2} \cdot \min\left\{1, \frac{\mu}{2LK}\right\}\right)^k.
\]

We thus recover the linear convergence result in [31, Theorem 1] with a sharper rate. For all other values of \(\nu, p, \gamma\) we have \(r = \frac{\gamma q}{p} < 1\). To simplify notation and exposition for this case, assume that \(\text{subopt}_0 \leq M\). Thus Theorem 3 yields the sublinear rate
\[
\text{subopt}_k \leq \left(1 - \frac{1}{2} \cdot \frac{\mu}{2LK}\right)^k \cdot \left(\frac{1}{M^{\frac{1}{\nu-1}}} + \frac{1-r}{q} \cdot \frac{k}{M^{\frac{1}{q-1}}}\right)^{\frac{\nu}{\nu-1}} \leq \left(\frac{(\nu + 1) \cdot M^{\frac{1}{\nu-1}}}{1 - \gamma(\nu + 1)/p} \cdot \frac{1}{k}\right)^{\frac{1-\gamma(\nu+1)/p}{p}}.
\]

Hence we extend the sublinear convergence results in [31, Theorem 1] and in [14, Theorem 2.9]. Both [31] and [14] consider only the special case \(\nu = 1\). In this special case we get \(r = 2\gamma/p\) and thus
\[
\text{subopt}_k \leq \left(\frac{2M}{1 - 2\gamma/p} \cdot \frac{1}{k}\right)^{\frac{1}{1-\gamma/p}}.
\]

In particular, we recover the \(O\left(1/k^{1-\gamma/p}\right)\) sublinear convergence rate in [31, Theorem 1] which only considers the case \(p = 2\). We also recover the \(O\left(1/k^{1-2\gamma/p}\right)\) sublinear
convergence result in [14, Theorem 2.9] which only considers the case \( \gamma \in [0, 1/2] \). Like both [14] and [31], we also recover the \( O(1/k^2) \) sublinear convergence result in [8, Theorem 2] which only considers the case \( p = 2 \) and \( \gamma = 1/2 \). In all cases we get slightly simpler and sharper constants.

5 Local growth property

We next describe an extension of the main developments in Section 2 to a local setting. The same extension applies to the developments in Section 4.

Consider the setting of Section 2. Suppose \( q > 1 \) and \( r \in [0, 1] \). We shall say that the pair \((D, \text{gap})\) satisfies the \((q, r)\) local growth property if there exist constants \( \epsilon > 0 \) and \( M > 0 \) such that for all \( x \in \text{dom}(\Psi) \), \( g := \nabla f(x) \), and \( s := \partial \Psi^*(-g) \) that satisfy \( \text{gap}(x, g) < \epsilon \) we have

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

(44)

It is straightforward to see that local versions of Theorem 1 and Theorem 2 also hold if we replace the growth property (10) with the local growth property (44) provided \( \text{gap}_0 = \text{gap}(x_0, g_0) < \epsilon \). It is also evident that the local growth property (44) is less stringent than the growth property (10). In particular, the natural local version of Proposition 4 holds if the uniform smoothness and uniform convexity assumptions hold on a neighborhood of an optimal solution to \( (1) \). The following proposition, which is a straightforward extension of some arguments in [14], describes another class of problems that satisfy the local growth condition (44). Proposition 6 provides a counterpart to Proposition 4(b) that relaxes the assumption of \( p \)-uniform convexity of \( C \) and boundedness away from zero of \( \nabla f \) on \( C \) to the far less stringent local scaling inequality (45) that only involves \( \nabla f(x^*) \) where \( x^* \in C \) is a minimizer of \( (1) \). This relaxed assumption comes at the expense of downgrading the growth condition for \( r = q/p \) to the local growth condition for \( r = q(q-1)/(p(p-1)) \). As noted in [14, Remark 2.6], in the special case when \( q = p = 2 \) there is no downgrade in the value of \( r \). For all other values of \( q \in (1, 2] \) and \( p \geq 2 \) the downgrade factor \( (q-1)/(p-1) \) on \( r \) is the price we pay for relaxing the uniform convexity and boundedness assumptions to the less stringent local scaling assumption.

**Proposition 6.** Suppose \( \Psi = \delta_C \) for a closed convex set \( C \subseteq \mathbb{R}^n \), \( x^* \in C \) is a minimizer of \( (1) \), \( \nu \in (0, 1] \), and \( p \geq 2 \). The local growth property (44) holds for \( \epsilon = 1 \), \( q = 1 + \nu \), and \( r = q(q-1)/(p(p-1)) \) if \( \nabla f \) is \( \nu \)-Hölder continuous on \( C \) and there exists \( \sigma > 0 \) such that the following local scaling inequality holds for all \( x \in C \)

\[
\langle \nabla f(x^*), x - x^* \rangle \geq \sigma \cdot \|x - x^*\|^p.
\]  

(45)

**Proof.** From (45) and the convexity of \( f \) it follows that for \( x \in C \) and \( g = \nabla f(x) \)

\[
\text{gap}(x, g) \geq f(x) - f(x^*) \geq \langle \nabla f(x^*), x^* - x \rangle \geq \sigma \cdot \|x - x^*\|^p.
\]  

(46)
In addition, for \( s = \arg\min_{y \in C} \langle g, y \rangle = \arg\min_{y \in C} \langle \nabla f(x), y \rangle \) the local scaling inequality (45) implies that
\[
\langle \nabla f(x^*) - \nabla f(x), s - x^* \rangle \geq \langle \nabla f(x^*), s - x^* \rangle \geq \sigma \cdot \| s - x^* \|^p.
\]
(47)

Thus the \( \nu \)-Hölder continuity of \( \nabla f \) and Hölder’s inequality imply that for some finite Hölder continuity constant \( L > 0 \)
\[
L \| x^* - x \|^\nu \cdot \| s - x^* \| \geq \sigma \cdot \| s - x^* \|^p.
\]
Hence \( \sigma \cdot \| s - x^* \|^{p-1} \leq L \| x^* - x \|^\nu \) and thus (46) yields
\[
\text{gap}(x, g) \geq \sigma \cdot \| x - x^* \|^p \geq \sigma \cdot \left( \frac{\sigma}{L} \right)^\frac{p}{p-1} \| s - x^* \|^\frac{p(p-1)}{\nu}.
\]
(48)

Let \( \epsilon := 1 \). From (46) and (48) it follows that for \( \text{gap}(x, g) < \epsilon \)
\[
\begin{align*}
\| s - x \| & \leq \| x^* - x \| + \| s - x^* \| \\
& \leq \frac{1}{\sigma^\frac{1}{p}} \cdot \text{gap}(x, g)^\frac{1}{p} + \frac{1}{\sigma^\frac{p}{p(p-1)}} \cdot \left( \frac{L}{\sigma} \right)^\frac{p}{p-1} \cdot \text{gap}(x, g)^\frac{p(p-1)}{\nu}
\end{align*}
\]
\[
\leq \left( \frac{1}{\sigma^\frac{1}{p}} + \frac{1}{\sigma^\frac{p}{p(p-1)}} \cdot \left( \frac{L}{\sigma} \right)^\frac{p}{p-1} \right) \cdot \text{gap}(x, g)^\frac{p}{p-1}.
\]
The last step above holds because \( \text{gap}(x, g) < \epsilon = 1 \) and \( \nu/(p-1) \in (0, 1] \) imply that \( \text{gap}(x, g)^\frac{p}{p-1} \leq \text{gap}(x, g)^\frac{\nu}{p(p-1)} \). Therefore (24) implies that for \( \text{gap}(x, g) < \epsilon \)
\[
\mathcal{D}(x, s, \theta) = \mathcal{D}_f(x + \theta(s - x), x)
\]
\[
\leq \frac{L \theta^{1+\nu}}{1+\nu} \cdot \| s - x \|^{1+\nu}
\]
\[
\leq L \cdot \left( \frac{1}{\sigma^\frac{1}{p}} + \frac{1}{\sigma^\frac{p}{p(p-1)}} \cdot \left( \frac{L}{\sigma} \right)^\frac{1}{p-1} \cdot \frac{\theta^{1+\nu}}{1+\nu} \right) \cdot \text{gap}(x, g)^\frac{\nu(1+\nu)}{p(p-1)}.
\]
Therefore whenever \( \text{gap}(x, g) < \epsilon \) the local growth property (44) holds for \( q = 1+\nu \) and \( r = q(q-1)/(p(p-1)) \) with
\[
M = L \cdot \left( \frac{1}{\sigma^\frac{1}{p}} + \frac{1}{\sigma^\frac{p}{p(p-1)}} \cdot \left( \frac{L}{\sigma} \right)^\frac{1}{p-1} \right)^{1+\nu}.
\]

\[
\square
\]

References

[1] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.
[2] 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.

[3] J. Bolte, A. Daniilidis, and A. Lewis. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[4] 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.

[5] K. Bredies, M. Carioni, S. Fanzon, and F. Romero. A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization. *Foundations of Computational Mathematics*, pages 1–66, 2022.

[6] K. Bredies, D. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009.

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

[8] D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2015.

[9] D. Garber, A. Kaplan, and S. Sabach. Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems. *Mathematical Programming*, 186(1):185–208, 2021.

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

[11] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015.

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

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

[14] 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.
[15] T. Kerdreux, A. d’Aspremont, and S. Pokutta. Restarting Frank–Wolfe: Faster rates under Hölderian error bounds. *Journal of Optimization Theory and Applications*, 192(3):799–829, 2022.

[16] 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.

[17] K. Kunisch and D. Walter. On fast convergence rates for generalized conditional gradient methods with backtracking stepsize. *Numerical Algebra, Control and Optimization*, pages 1–29, 2022.

[18] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l’institut Fourier*, volume 48, pages 769–783, 1998.

[19] E. Levitin and B. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[20] H. Li and L. Ding. Generalized conditional gradient method for elastic-net regularization. *Journal of Computational and Applied Mathematics*, 403:113872, 2022.

[21] S. Lojasiewicz. Ensembles semi-analytiques. *Lectures Notes IHES (Bures-sur-Yvette)*, 1965.

[22] 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.

[23] 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.

[24] H. Mine and M. Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *Journal of Optimization Theory and Applications*, 33:9–23, 1981.

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

[26] J. Nocedal and S. Wright. *Numerical Optimization, Second Edition*. Springer, 2006.

[27] F. Rinaldi and D. Zeffiro. A unifying framework for the analysis of projection-free first-order methods under a sufficient slope condition. *arXiv preprint arXiv:2008.09781*, 2020.

[28] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, pages 1–30, 2018.
[29] Q. Van Nguyen. Variable quasi-Bregman monotone sequences. *Numerical Algorithms*, 73(4):1107–1130, 2016.

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

[31] Y. Xu and T. Yang. Frank-Wolfe method is automatically adaptive to error bound condition. *arXiv preprint arXiv:1810.04765*, 2018.

[32] Y. Yu, X. Zhang, and D. Schuurmans. Generalized conditional gradient for sparse estimation. *Journal of Machine Learning Research*, 18:1–46, 2017.