General Hölder Smooth Convergence Rates Follow from Specialized Rates Assuming Growth Bounds

Benjamin Grimmer

Received: 26 April 2021 / Accepted: 1 February 2023 / Published online: 25 February 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Often in the analysis of first-order methods for both smooth and nonsmooth optimization, assuming the existence of a growth/error bound or KL condition facilitates much stronger convergence analysis. Hence, separate analysis is typically needed for the general case and for the growth bounded cases. We give meta-theorems for deriving general convergence rates from those assuming a growth lower bound. Applying this simple but conceptually powerful tool to the proximal point, subgradient, bundle, dual averaging, gradient descent, Frank–Wolfe and universal accelerated methods immediately recovers their known convergence rates for general convex optimization problems from their specialized rates. New convergence results follow for bundle methods, dual averaging and Frank–Wolfe. Our results can lift any rate based on Hölder continuous gradients and Hölder growth bounds. Moreover, our theory provides simple proofs of optimal convergence lower bounds under Hölder growth from textbook examples without growth bounds.

Keywords First-order methods · Convex optimization · Convergence rates · Growth/error bounds · Restarting schemes

Mathematics Subject Classification 65K05 · 90C25 · 90C30

1 Introduction
Recent interest in first-order methods for convex optimization has grown out of their successful application in high-dimensional statistics, signal processing and data science. The effectiveness of first-order methods in solving these large-scale applications
is fundamentally tied to the problem’s underlying geometry. This has led to the development of robust convergence theory, describing the algorithmic speed-ups induced by the existence of different problem structures.

Typically first-order optimization methods rely on the objective function possessing some continuity or smoothness structure. In nonsmooth optimization, an assumption like the objective function being uniformly Lipschitz continuous is usually needed. In smooth optimization, an assumption like the objective’s gradient being uniformly Lipschitz is often made. In this work, we consider Hölder smooth objective functions (see (2)), which captures both of these standard settings and facilitates the development of universal or blackbox methods as done in the recent works [5, 9, 13, 25, 29].

Throughout the first-order optimization literature, improved convergence rates typically follow from assuming the given objective function satisfies a growth/error bound or Kurdyka–Łojasiewicz condition (see [1, 2, 12, 19, 20] for a sample of works developing these ideas). We consider the formalization of this through Hölder growth bounds (see (3)) with respect to some growth exponent $p$. The two most important cases of this bound are Quadratic Growth given by $p = 2$ (generalizing strong convexity [21]) and Sharp Growth given by $p = 1$ (which occurs broadly in nonsmooth optimization [3]). Beyond these two special cases, nearly every objective function of interest is semialgebraic, from which the results of [1] ensure a growth bound holds locally with some exponent (although it may be large and thus correspond to quite flat growth). Considering the generalization given by the Kurdyka–Łojasiewicz property [2], prior works derived dimension-dependent bounds on this exponent [16] and calculus relating them to those of underlying component functions [17].

Different convergence proofs are typically employed for minimization with or without assuming the existence of a given growth lower bound. Table 1 summarizes the number of iterations required to guarantee $\epsilon$-accuracy for several well-known first-order methods (see Sect. 2 for a precise description of these algorithms and “Appendix,” for example, details applying our theory).

**Our Contribution.** This work presents a pair of meta-theorems for deriving general convergence rates from rates that assume the existence of a growth lower bound. In terms of Table 1, we show that each convergence rate implies all of the convergence rates to its left: The quadratic growth column’s rates imply all of the general setting’s rates and each sharp growth rate implies that method’s general and quadratic growth rate. More generally, our results show that any convergence rate assuming growth with exponent $p$ implies rates for any growth exponent $q > p$ and for the general setting.

Interestingly, this recovers the proximal point method’s superlinear convergence when $1 < p < 2$ from its finite guarantee when $p = 1$. Moreover, new convergence guarantees follow dual averaging and bundle methods, which previously only had theory under quadratic growth\(^1\). Inspired by these results, the subsequent work [5] gave direct (and nontrivial) derivations of the bundle method rates yielded by our theory. For the Frank–Wolfe variant [9, Algorithm 2] when $p > 2$, lifting their $p = 2$ rate gives a faster rate of $O(1/\alpha^{2/p} \epsilon^{1-2/p})$ in terms of $\epsilon$ than the previously best known rate of $O(1/\alpha^{1/p-1} \epsilon^{p-1})$ proved by [9] (see “Appendix E” for further discussion).

\(^1\) [4, Theorem 3] assumed the stronger condition of strong convexity, but only quadratic growth is used in its proof, being invoked just after equation (38) therein.
| Method                                      | General | $p > 2$ | Quad. Growth | $1 < p < 2$ | Sharp Growth |
|---------------------------------------------|---------|---------|--------------|-------------|--------------|
| Proximal point method [7, 30]              | $O\left(\frac{1}{\epsilon}\right)$ | $O\left(\frac{1}{\alpha^{2/p} \epsilon^{1-2/p}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\alpha}\right)$ | $O\left(\frac{\log \log(\Delta/\epsilon)}{\alpha^{2/p}}\right)$ | $O\left(\frac{\Delta - \epsilon}{\alpha^2}\right)$ |
| Projected subgradient method [27, 28]      | $O\left(\frac{1}{\epsilon^2}\right)$ | $O\left(\frac{1}{\alpha^{2/p} \epsilon^{2-2/p}}\right)$ | $O\left(\frac{1}{\epsilon^{2/p}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\alpha^{2/p}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\alpha^2}\right)$ |
| Dual averaging variants [4, 24]             | $O\left(\frac{1}{\epsilon^2}\right)$ | $O\left(\frac{1}{\alpha^{2/p} \epsilon^{2-2/p}}\right)$ | $O\left(\frac{1}{\epsilon^{2/p}}\right)$ | None Known | None Known |
| Proximal Bundle Method [5, 6, 11]          | $O\left(\frac{1}{\epsilon^3}\right)$ | $O\left(\frac{1}{\alpha^{2/3} \epsilon^{2-2/3}}\right)$ | $O\left(\frac{1}{\epsilon^{2/3}}\right)$ | $O\left(\frac{1}{\epsilon} + \frac{1}{\alpha^{4/3} \epsilon^{3-4/3}}\right)$ | $O\left(\frac{1}{\epsilon} + \frac{\Delta}{\alpha^4}\right)$ |
| Projected gradient descent [23]            | $O\left(\frac{1}{\epsilon}\right)$ | $O\left(\frac{1}{\alpha^{2/p} \epsilon^{1-2/p}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\alpha}\right)$ | $-$ | $-$ |
| Frank–Wolfe variants [8, 9]                | $O\left(\frac{1}{\epsilon}\right)$ | $O\left(\frac{1}{\alpha^{2/p} \epsilon^{1-2/p}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\alpha}\right)$ | $-$ | $-$ |
| Restarted universal method [29, 31]        | $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ | $O\left(\frac{1}{\alpha^{1/3} \epsilon^{1-2/3}}\right)$ | $O\left(\frac{\log(\Delta/\epsilon)}{\sqrt{\alpha}}\right)$ | $-$ | $-$ |

The proximal point method makes no smoothness or continuity assumptions. The subgradient, dual averaging and bundle method rates assume Lipschitz continuity, and the gradient descent, Frank–Wolfe and universal method rates assume Lipschitz gradient. Note several of these methods have been analyzed beyond these Lipschitz/smooth settings.
A natural question is whether the reverse implications hold (whether each column of Table 1 implies the rates to its right). If one is willing to modify the given first-order method, then the literature already provides a partial answer through restarting schemes [18, 22, 29, 31, 34]. Such schemes repeatedly run a given first-order method until some criterion is met and then restart the method at the current iterate. For example, Renegar and Grimmer [29] show how a general convergence rate can be extended to yield a convergence rate under Hölder growth. Then together with our theory, any rate with or without a growth bound gives rates for every growth exponent and the general case.

Outline. Section 2 first formalizes our model for first-order methods and then states our rate lifting theorems and their restarting extensions. Section 3 proves our main results via analysis based on related Fenchel conjugates. Details for the algorithms and applications shown in Table 1 are deferred to “Appendix.”

2 Rate Lifting Theorems

We are primarily interested in minimizing a closed proper convex function \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \) over a closed convex set \( Q \subseteq \mathbb{R}^n \)

\[
\min_{x \in Q} f(x)
\]

that attains its minimum value of \( f_* \) on some set \( X_* \subseteq Q \).

Most first-order methods for minimizing such problems rely on \( f \) being \((L, \eta)\)-Hölder smooth, defined for some \( L \geq 0 \) and \( 0 \leq \eta \leq 1 \) as

\[
\|g - g'\| \leq L\|x - x'\|^\eta \quad \text{for all } x, x' \in Q, g \in \partial f(x), g' \in \partial f(x') ,
\]

where \( \partial f(x) = \{g \in \mathbb{R}^n \mid f(x') \geq f(x) + \langle g, x' - x \rangle \forall x' \in \mathbb{R}^n \} \) is the subdifferential of \( f \) at \( x \). When \( \eta = 1 \), this corresponds to the common assumption of \( L \)-smoothness (i.e., having \( \nabla f(x) \) be \( L \)-Lipschitz). When \( \eta = 0 \), this captures the standard nonsmooth optimization model of having \( f(x) \) itself be \( L \)-Lipschitz. The analysis of many first-order methods further assumes \((\alpha, p)\)-Hölderian growth (or a similar KL/error bound condition), defined by

\[
f(x) \geq f_* + \alpha \text{dist}(x, X_*)^p \quad \text{for all } x \in Q ,
\]

where \( \text{dist}(x, X_*) = \inf\{\|x - x_*\| \mid x_* \in X_*\} \). When \( p = 2 \), this corresponds to quadratic growth, capturing the common setting of strongly convex objectives, and when \( p = 1 \), this corresponds to sharp growth.

Recall Table 1 shows the improved convergence rates of several first-order methods relying on Hölder smoothness and Hölder growth bounds. Namely, it considers the proximal point method with any step size \( \rho > 0 \) defined by
\[ x_{k+1} = \text{prox}_{\rho, f}(x_k) := \arg\min_{x \in Q} \left\{ f(x) + \frac{1}{2\rho} \| x - x_k \|^2 \right\}, \quad (4) \]
the projected subgradient method defined by
\[ x_{k+1} = \text{proj}_Q(x_k - \rho_k g_k) \text{ for some } g_k \in \partial f(x_k), \quad (5) \]
where \( \text{proj}_Q(\cdot) \) denotes orthogonal projection onto \( Q \), and projected gradient descent
\[ x_{k+1} = \text{proj}_Q(x_k - \rho_k \nabla f(x_k)) \quad (6) \]
with step size \( \rho_k = \| \nabla f(x_k) \|^{(1-\eta)/\eta} / L^{1/\eta} \). In “Appendix,” we apply our results to the existing convergence theory for these and the more sophisticated proximal bundle method\(^2\) of [5, 14, 33], Frank–Wolfe/conditional gradient methods [8, 9] and a restarted variant of the universal gradient method of [25]. Note the references above analyzed the universal settings of Hölder smoothness and Hölder growth for these three of these methods.

2.1 Formalization of Rate Lifting Theorems

Now we formalize our model for a generic first-order method \( \text{fom} \) to which our rate lifting theorems will apply. Note that for it to be meaningful to lift a convergence rate to apply to general problems, the inputs to \( \text{fom} \) need to be independent of the existence of a growth bound (3), but may depend on the constraints \( Q \), Hölder smoothness constants \( (L, \eta) \), or optimal objective value \( f_* \). We make the following three assumptions about \( \text{fom} \):

(A1) The method \( \text{fom} \) computes a sequence of iterates \( \{x_k\}_{k=0}^{\infty} \) in \( Q \). The next iterate \( x_{k+1} \) is determined by the first-order oracle values \( \{(f(x_j), g_j)\}_{j=0}^{k+1} \), where \( g_j \in \partial f(x_j) \).

(A2) The distance from any \( x_k \) to some fixed \( x_* \in X_* \) is at most some constant \( D > 0 \).

(A3) For some \( p \geq 1 \), there exists a function \( K: \mathbb{R}^3_{++} \to \mathbb{R} \) such that if \( f \) is \( (L, \eta) \)-Hölder smooth and possesses \( (\alpha, p) \)-Hölder growth on \( B(x_*, D) \cap Q \), then for any \( \epsilon > 0 \), \( \text{fom} \) finds an \( \epsilon \)-minimizer \( x_k \) with
\[ k \leq K(f(x_0) - f_*, \epsilon, \alpha). \]

On the Generality of (A1)–(A3). Note that (A1) allows the computation of \( x_{k+1} \) to depend on the function and subgradient values at \( x_{k+1} \). Hence, the proximal point method is included in our model since it is equivalent to \( x_{k+1} = x_k - \rho_k g_{k+1} \), where \( g_{k+1} \in \partial f(x_{k+1}) \). We remark that (A2) holds for all of the previously mentioned algorithms under reasonable selection of their step size parameters. In fact, many common first-order methods are nonexpansive, giving \( D = \| x_0 - x_* \| \).

\(^2\) Stronger convergence rates have been established for related level bundle methods [10, 13, 15], which share many core elements with proximal bundle methods.
Alternatively, if the constraint set $Q$ is compact, (A2) trivially holds. Lastly, note that the convergence bound $K(f(x_0) - f^*, \epsilon, \alpha)$ in (A3) can depend on $L$, $\eta$, $p$ even though our notation does not enumerate this. If one wants to make no continuity or smoothness assumptions about $f$, $(L, \eta)$ can be set as $(\infty, 0)$. (The proximal point method is one such example as it converges independent of such structure.)

The following pair of convergence rate lifting theorems show that these assumptions suffice to give general convergence guarantees without Hölder growth and guarantees for Hölder growth with any exponent $q > p$. In terms of Table 1, these show that each column implies the rates to its left.

**Theorem 2.1** Consider any method $\text{fom}$ satisfying (A1)–(A3) and any $(L, \eta)$-Hölder smooth function $f$. For any $\epsilon > 0$, $\text{fom}$ will find $f(x'_k) - f^* \leq \epsilon$ by iteration

$$k \leq K(f(x_0) - f^*, \epsilon, \epsilon/D^p),$$

where $x'_k \in \arg\min_{x \in Q} \{f(x_k) + \langle g_k, x-x_k \rangle + \frac{L}{\eta+1} \|x-x_k\|^{\eta+1} \}$ minimizes a Hölder smooth first-order model of $f$ over $Q$.

**Theorem 2.2** Consider any method $\text{fom}$ satisfying (A1)–(A3) and any $(L, \eta)$-Hölder smooth function $f$ possessing $(\alpha, q)$-Hölder growth with $q > p$. For any $\epsilon > 0$, $\text{fom}$ will find $f(x'_k) - f^* \leq \epsilon$ by iteration

$$k \leq K(f(x_0) - f^*, \epsilon, \alpha^{p/q} \epsilon^{1-p/q}),$$

where $x'_k \in \arg\min_{x \in Q} \{f(x_k) + \langle g_k, x-x_k \rangle + \frac{L}{\eta+1} \|x-x_k\|^{\eta+1} \}$ minimizes a Hölder smooth first-order model of $f$ over $Q$.

Note these guarantees apply to the auxiliary point $x'_k$. For nonsmooth settings, when $\eta = 0$, this is exactly $\text{fom}$‘s iterates $x_k = x'_k$. When $\eta \in (0, 1]$, this point is one projected gradient descent step away $x'_k = \text{proj}_Q(x_k - \alpha_k \nabla f(x_k))$ with $\alpha_k = \|x_k - x'_k\|^{1-\eta}/L$. Although $\alpha_k$ is implicitly defined based on $x'_k$, it can often be computed with minor care: When $\eta = 1$, $\alpha_k = 1/L$. When $Q = \mathbb{R}^n$, $\alpha_k = \|\nabla f(x_k)\|^{(1-n)/n} / L^{1/n}$. Otherwise, a line-searching procedure can be applied to find $\alpha_k$. One shortcoming of our lifted guarantees is the need to compute an orthogonal projection to compute $x'_k$. For methods utilizing projections, this cost is marginal, but for methods like Frank–Wolfe, additional computation is needed (perhaps just inexacty computing just the final $x'_k$).

Applying these rate lifting theorems amounts to simply substituting $\alpha$ with $\epsilon/D^p$ or $\alpha^{p/q} \epsilon^{1-p/q}$ in any guarantee depending on $(\alpha, p)$-Hölder growth. For example, the projected subgradient method using the Polyak step size satisfies (A2) with $D = \|x_0 - x_*\|$ and converges for any $L$-Lipschitz objective with quadratic growth $p = 2$ at rate

$$K(f(x_0) - f^*, \epsilon, \alpha) = \frac{8L^2}{\alpha \epsilon}$$
establishing (A3) (a proof of this is given in “Appendix” for completeness). Then applying Theorem 2.1 recovers the method’s classic convergence rate as

$$K(f(x_0) - f^*, \varepsilon, \varepsilon / D^2) = \frac{8L^2 \|x_0 - x^*\|^2}{\varepsilon^2}.$$  

For convergence rates that depend on $f(x_0) - f^*$, this simple substitution falls short of recovering the method’s known rates, often off by a log term. Again taking the subgradient method as an example, under sharp growth $p = 1$, convergence occurs at a rate of

$$K(f(x_0) - f^*, \varepsilon) = \frac{4L^2}{\alpha^2} \log_2 \left( \frac{f(x_0) - f^*}{\varepsilon} \right).$$

Then Theorem 2.1 ensures the following weaker general rate

$$K(f(x_0) - f^*, \varepsilon, \varepsilon / D) = \frac{4L^2 \|x_0 - x^*\|^2}{\varepsilon^2} \log_2 \left( \frac{f(x_0) - f^*}{\varepsilon} \right)$$

and Theorem 2.2 ensures the weaker quadratic growth rate

$$K(f(x_0) - f^*, \varepsilon, \varepsilon / D) = \frac{4L^2}{\alpha \varepsilon} \log_2 \left( \frac{f(x_0) - f^*}{\varepsilon} \right).$$

In the following subsection, we provide corollaries that remedy this issue.

### 2.2 Improved Rate Lifting Theorems via Restarting

The ideas from restarting schemes can also be applied to our rate lifting theory. We consider the following conceptual restarting method restarting-fom that repeatedly halves the objective gap: Set an initial target accuracy of $\tilde{\varepsilon} = 2^{N-1} \varepsilon$ with $N = \lceil \log_2((f(x_0) - f^*)/\varepsilon) \rceil$. Iteratively run fom until an $\tilde{\varepsilon}$-optimal solution is found, satisfying

$$f(x'_k) - f^* \leq \tilde{\varepsilon}$$

for $x'_k \in \arg\min_{x \in Q} \{f(x_k) + \langle g_k, x - x_k \rangle + \frac{L}{\eta+1} \|x - x_k\|^{\eta+1} \}$ and then restart fom at the point $x_0 \leftarrow x'_k$ with new target accuracy $\tilde{\varepsilon} \leftarrow \tilde{\varepsilon} / 2$.

Note for the proximal point method (4), subgradient method with Polyak’s step size (5) and gradient descent (6), this restarting will not change the algorithm’s trajectory: This follows as (i) all three of these methods have the iterates $\{y_k\}$ produced by fom initialized at $y_0 = x_T$ satisfy $y_k = x_{k+T}$ and (ii) the proximal point method and subgradient method both have $x'_k = x_k$ and projected gradient descent has $x'_k = x_{k+1}$. As a result, the following corollaries of our lifting theorems apply directly to these three methods without restarting.
Corollary 2.1  Consider any method $\mathcal{fom}$ satisfying (A1)–(A3) and any $(L, \eta)$-Hölder smooth function $f$. For any $\epsilon > 0$, restart-$\mathcal{fom}$ will find $f(x'_k) - f_* \leq \epsilon$ after at most

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, 2^n\epsilon/D^p)$$

total iterations.

Proof  Observe that restart-$\mathcal{fom}$ must have found an $\epsilon$-minimizer after the $N$th restart. Our corollary then follows from bounding the number of iterations required for each of these $N$ restarts. Run $i \in \{1 \ldots N\}$ of $\mathcal{fom}$ has initial objective gap at most $2^{N+1-i}\epsilon$, and so, Theorem 2.1 ensures an $2^{N-i}\epsilon$-minimizer is found after at most

$$K(2^{N+1-i}\epsilon, 2^{N-i}\epsilon, 2^{N-i}\epsilon/D^p)$$

iterations. Summing this bound over all $i$ gives the claimed result. \hfill $\square$

Corollary 2.2  Consider any method $\mathcal{fom}$ satisfying (A1)–(A3) and any $(L, \eta)$-Hölder smooth function $f$ possessing $(\alpha, q)$-Hölder growth with $q > p$. For any $\epsilon > 0$, restart-$\mathcal{fom}$ will find $f(x'_k) - f_* \leq \epsilon$ after at most

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, \alpha p/q (2^n\epsilon)^{1-p/q})$$

total iterations.

Proof  Along the same lines as the proof of Corollary 2.1, this corollary follows from bounding the number of iterations required for each of restart-$\mathcal{fom}$’s $N$ restarts. Run $i \in \{1 \ldots N\}$ of $\mathcal{fom}$ has initial objective gap at most $2^{N+1-i}\epsilon$, and so, Theorem 2.2 ensures an $2^{N-i}\epsilon$-minimizer is found after at most

$$K(2^{N+1-i}\epsilon, 2^{N-i}\epsilon, \alpha p/q (2^{N-i}\epsilon)^{1-p/q})$$

iterations. Summing this bound over all $i$ gives the claimed result. \hfill $\square$

Applying these corollaries to every entry in Table 1 verifies our claim that each column implies those to its left. For example, the proximal point method has (A2) hold with $D = \|x_0 - x_*\|$ and (A3) hold with

$$K(f(x_0) - f_*, \epsilon, \alpha) = \frac{f(x_0) - f_* - \epsilon}{\rho \alpha^2}$$

under sharp growth $p = 1$. (A proof of this fact is given in “Appendix” for completeness.) Then Corollary 2.1 recovers the proximal point method’s general convergence rate of $\frac{1}{\rho \alpha^2}$. Springer
\[
\sum_{n=0}^{N} K(2^{n+1} \epsilon, 2^n \epsilon, 2^{n} \epsilon/D) = \sum_{n=0}^{N} \frac{2^{n+1} \epsilon - 2^n \epsilon}{\rho(2^n \epsilon/D)^2} = \sum_{n=0}^{N} \frac{D^2}{\rho 2^n \epsilon} = \frac{2\|x_0 - x_\star\|^2}{\rho \epsilon}
\]

and Corollary 2.2 recovers its linear convergence rate under quadratic growth \(q = 2\) of

\[
\sum_{n=0}^{N} K(2^{n+1} \epsilon, 2^n \epsilon, \alpha^{1/2} (2^n \epsilon)^{1/2}) = \sum_{n=0}^{N} \frac{2^{n+1} \epsilon - 2^n \epsilon}{\rho (2^n \epsilon^{1/2} \alpha^{1/2} \epsilon^{1/2})^2} = \sum_{n=0}^{N} \frac{1}{\rho \alpha} = \frac{N}{\rho \alpha}.
\]

Likewise, the \(O(\sqrt{L/\alpha} \log(f(x_0) - x(x_\star))/\epsilon)\) rate that follows from restarting an accelerated method for \(L\)-smooth optimization under quadratic growth \(p = 2\) recovers Nesterov’s classic accelerated convergence rate as

\[
\sum_{n=0}^{N} K(2^{n+1} \epsilon, 2^n \epsilon, 2^n \epsilon/D^2) = \sum_{n=0}^{N} O\left(\sqrt{L D^2/2^n \epsilon \log(2)}\right) = O\left(\sqrt{L D^2/\epsilon}\right).
\]

### 2.3 Recovering Lower Bounds on Oracle Complexity

The contrapositive of our rate lifting theorems immediately lifts complexity lower bounds from the general case to apply to the specialized case of Hölder growth. Well-known simple examples [23] give complexity lower bounds when no growth bound is assumed for first-order methods satisfying

(A4) For every \(k \geq 0\), \(x_{k+1}\) lies in the span of \(\{g_i\}_{i=0}^k\).

Optimal bounds under Hölder growth were claimed by Nemirovski and Nesterov [22, page 26], although no proof is given. Here we note that the simple examples from the general case combined with our theory suffice to give concise derivations of lower bounds on the convergence rates under growth conditions.

For \(M\)-Lipschitz nonsmooth optimization, a simple example shows any method satisfying (A4) cannot guarantee finding an \(\epsilon\)-minimizer in fewer than \(M^2\|x_0 - x_\star\|^2/16\epsilon^2\) subgradient evaluations. Consequently, applying Theorem 2.1, any method satisfying (A1)-(A4) for \(M\)-Lipschitz optimization with \((\alpha, p)\)-Hölder growth must have its rate \(K(f(x_0) - f_\star, \epsilon, \alpha)\) bounded below by

\[
K(f(x_0) - f_\star, \epsilon, \epsilon/D^p) \geq \frac{M^2\|x_0 - x_\star\|^2}{16\epsilon^2}.
\]

For example, this ensures the Polyak subgradient method’s \(O(1/\alpha \epsilon)\) rate under quadratic growth cannot be improved by more than small constants. The exponent on \(\alpha\) cannot decrease since that would beat the general cases’ lower bound dependence on \(\|x_0 - x_\star\|^2\) (as \(D = \|x_0 - x_\star\|\) here), and similarly, the exponent of \(\epsilon\) cannot be improved due to the lower bound dependence of \(1/\epsilon^2\).

Likewise, for \(L\)-smooth optimization, a simple example shows no method satisfying (A4) can guarantee computing an \(\epsilon\)-minimizer using fewer than \(\sqrt{3L\|x_0 - x_\star\|^2}/32\epsilon\) subgradient evaluations.
iterations. Applying Theorem 2.1, we conclude any method satisfying (A1)-(A4) for smooth optimization with \((\alpha, p)\)-Hölder growth has its rate \(K(f(x_0) - f^*, \epsilon, \alpha)\) bounded below by

\[
K(f(x_0) - f^*, \epsilon, \epsilon/D^p) \geq \sqrt{3L\|x_0 - x^\star\|^2 / 32\epsilon}.
\]

One can immediately conclude the optimal convergence smooth strongly convex functions grows with at least \(O(1/\sqrt{\alpha})\). (Otherwise lifting the rate would violate the general lower bounds dependence of \(O(\|x_0 - x^\star\|)\).)

### 3 Proofs of the Rate Lifting Theorems 2.1 and 2.2

The proofs of our two main results (Theorems 2.1 and 2.2) rely on several properties of Fenchel conjugates, which we will first review. The Fenchel conjugate of a function \(f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}\) is

\[
f^\star(g) := \sup_x \{\langle g, x \rangle - f(x)\}.
\]

We say that \(f \geq h\) if for all \(x \in \mathbb{R}^n\), \(f(x) \geq h(x)\). The conjugate reverses this partial ordering, having \(f \geq h \implies h^\star \geq f^\star\). Applying the conjugate twice \(f^{**}\) gives the largest closed convex function majorized by \(f\) (that is, the “convex envelope” of \(f\)), and consequently, for closed convex functions, \(f^{**} = f\).

The \((L, q)\)-Hölder smoothness condition (2) is equivalent to having upper bounds of the following form hold for every \(x \in \mathbb{R}^n\) and \(g \in \partial f(x)\)

\[
f(x') \leq f(x) + \langle g, x' - x \rangle + \frac{L}{\eta + 1} \|x' - x\|^{\eta+1}.
\]

Taking the conjugate of this convex upper bound gives an equivalent dual condition: a function \(f\) is \((L, \eta)\)-Hölder smooth if and only if its Fenchel conjugate has the following lower bound for every \(g \in \mathbb{R}^n\) and \(x \in \partial f^\star(g)\)

\[
f^\star(g') \geq f^\star(g) + \langle x, g' - g \rangle + \begin{cases} 
\eta/(\eta+1)L^{1/\eta} \|g' - g\|^{(\eta+1)/\eta} & \text{if } \eta > 0 \\
\delta_{\|g' - g\| \leq L}(g') & \text{if } \eta = 0
\end{cases},
\]

where \(\delta_{\|g' - g\| \leq L}(g') = \begin{cases} 
0 & \text{if } \|g' - g\| \leq L \\
\infty & \text{otherwise}
\end{cases}\) is an indicator function. This corresponds to the Fenchel conjugate being sufficiently uniformly convex. (In particular when \(\eta = 1\), this dual condition is exactly strong convexity.)
3.1 Proof of Theorem 2.1

Suppose that no iteration \( k \leq T \) has \( x_k' \) as an \( \epsilon \)-minimizer of \( f \) over \( Q \). We consider the following convex auxiliary functions: At each \( x_k \), the \((L, \eta)\)-Hölder smoothness and subgradient \( g(x_k) = g_k \in \partial f(x_k) \) define

\[
h_{x_k}(x) := f(x_k) + \langle g(x_k), x - x_k \rangle + \frac{L}{\eta + 1} \| x - x_k \|^{\eta + 1}
\]

and at each minimizer \( \bar{x} \in X_* \) with subgradient \( g(\bar{x}) \in \partial f(\bar{x}) \) normal to the constraint set \(-g(\bar{x}) \in N_Q(\bar{x})\) (certifying the optimality of \( \bar{x} \)) define

\[
h_{\bar{x}}(x) := f_\ast + \langle g(\bar{x}), x - \bar{x} \rangle + \frac{L}{\eta + 1} \| x - \bar{x} \|^{\eta + 1}.
\]

Then we focus on the convex envelope surrounding these models

\[
h(x) := (\min \{ h_{x} (\cdot) | x \in \{ x_0, \ldots, x_T \} \cup X_* \})^{**} (x) .
\]

We consider the auxiliary minimization problem of \( \min_{x \in Q} h(x) \), which shares and improves on the structure of \( f \) as described in the following three lemmas.

**Lemma 3.1** The objectives \( f \) and \( h \) have \( f(\bar{x}) = h(\bar{x}) \) and \( g(\bar{x}) \in \partial h(\bar{x}) \) for each \( \bar{x} \in \{ x_0, \ldots, x_T \} \cup X_* \).

**Lemma 3.2** The objective \( h \) is \((L, \eta)\)-Hölder smooth.

**Lemma 3.3** The objective \( h \) has \((\epsilon / D^p, p)\)-Hölder growth on \( B(x_\ast, D) \cap Q \).

Before proving these three results, we show that they suffice to complete our proof. Lemma 3.1 and assumption (A1) together ensure that applying \( \min \) to either \( f \) or \( h \) produces the same sequence of iterates up to iteration \( T \). Since \( f \) and \( h \) both minimize on \( X_* \) (as each \( \bar{x} \in X_* \) has a subgradient \( g(\bar{x}) \in -N_Q(\bar{x}) \) for both \( f \) and \( h \)), no \( x_k \) with \( k \leq T \) is an \( \epsilon \)-minimizer of \( h \). Hence, applying the structural conditions from Lemmas 3.2 and 3.3 with assumption (A3) on \( h \) completes our rate lifting argument as

\[
T < K ( h(x_0) - h(x_\ast), \epsilon, \epsilon / D^p ) = K ( f(x_0) - f_\ast, \epsilon, \epsilon / D^p ) .
\]

3.1.1 Proof of Lemma 3.1

By definition, each \( g(x_i) \) provides a linear lower bound on \( f \) as

\[
f(x) \geq f(x_i) + \langle g(x_i), x - x_i \rangle .
\]

Noting that all \( \bar{x} \in \{ x_0, \ldots, x_T \} \cup X_* \) have \( h_x \geq f \), it follows that for all \( x \in Q \)

\[
\min \{ h_{\bar{x}}(x) | \bar{x} \in \{ x_0, \ldots, x_T \} \cup X_* \} \geq f(\bar{x}) + \langle g(\bar{x}), x - \bar{x} \rangle .
\]

 Springer
Since this linear lower bound is convex, it is also a lower bound on the convex envelope \( h \). Setting \( x = \bar{x} \), equality holds with this linear lower bound. Thus, \( f(\bar{x}) = h(\bar{x}) \) and \( g(\bar{x}) \) is also a subgradient of \( h \) at \( \bar{x} \).

### 3.1.2 Proof of Lemma 3.2

Here our proof relies on the dual perspective of \((L, \eta)\)-Hölder smoothness given by (8). Since each \( h_{\bar{x}} \) is \((L, \eta)\)-Hölder smooth on \( \mathbb{R}^n \), each \( h^*_{\bar{x}} \) satisfies the dual lower bounding condition (8). Observe that

\[
\begin{align*}
    h^*(y) &= \left( \min_{\tilde{x} \in \{x_0, \ldots, x_T\} \cup X_*} \{h_{\bar{x}}(\cdot)\} \right)^*(y) \\
    &= \sup_x \left\{ \langle y, x \rangle - \min_{\tilde{x} \in \{x_0, \ldots, x_T\} \cup X_*} \{h_{\bar{x}}(x)\} \right\} \\
    &= \max_{\tilde{x} \in \{x_0, \ldots, x_T\} \cup X_*} \{h_{\bar{x}}^*(y)\}.
\end{align*}
\]

Consequently, \( h^* \) also satisfies the dual condition (8) since it is the maximum of functions satisfying this lower bound. Therefore, \( h^{**} = h \) retains the \((L, \eta)\)-Hölder smoothness of each of the models \( h_{\bar{x}} \) and the original objective \( f \).

### 3.1.3 Proof of Lemma 3.3

For any \( x \in B(x_*, D) \in Q \) and \( k \in \{0, \ldots, T\} \), the function \( h_{x_k} \) lies above our claimed growth bound as

\[
h_{x_k}(x) \geq h_{x_k}(x') \geq f_* + \epsilon \geq f_* + \frac{\epsilon}{D^p} \|x - x_*\|^p,
\]

where the first inequality uses that \( x'_k \) minimizes \( h_{\bar{x}} \) over \( Q \), the second uses \( h_{\bar{x}} \geq f \) and no \( \epsilon \)-minimizer has been found, and the last inequality uses that \( x \in B(x_*, D) \). Similarly, our growth bound holds for any \( h_{\bar{x}} \) with \( \bar{x} \in X^* \) as

\[
h_{\bar{x}}(x) = f_* + \langle g(\bar{x}), x - \bar{x} \rangle + \frac{L}{\eta + 1} \|x - \bar{x}\|^{\eta + 1}
\]

\[
\geq f_* + \frac{L}{\eta + 1} \|x - \bar{x}\|^{\eta + 1}
\]

\[
\geq f_* + \frac{L}{(\eta + 1)D^{p-(\eta+1)}} \text{dist}(x, X_*)^p
\]

\[
\geq f_* + \frac{\epsilon}{D^p} \text{dist}(x, X_*)^p,
\]

where the first inequality uses the optimality condition that \(-g(\bar{x})\) is normal to \( Q \) and \( x \in Q \), the second inequality uses that \( x \in B(x_*, D) \), and the last inequality uses that \( \frac{L D^{p+1}}{\eta+1} \geq f(x_0) - f_* \geq \epsilon \). Hence, the minimum \( \min_{x \in \{x_0, \ldots, x_T\} \cup X_*} \{h_{\bar{x}}(x)\} \) satisfies

\( \geq \) Springer
our claimed Hölder growth lower bound. Since this lower bound is convex, its convex envelope $h$ must also satisfy the Hölder bound (3).

3.2 Proof of Theorem 2.2

Suppose that no iteration $k \leq T$ has $x'_k$ as an $\epsilon$-minimizer of $f$ over $Q$. Consider the auxiliary objective $h(x)$ defined by (9). Then Lemmas 3.1 and 3.2 show $h$ agrees with $f$ everywhere $\text{fom}$ visits and is $(L, \eta)$-Hölder smooth. From this, (A1) ensures that applying $\text{fom}$ to either $f$ or $h$ produces the same sequence of iterates up to iteration $T$. Since $f$ and $h$ both minimize on $X_*$, no $x_k$ with $k \leq T$ is an $\epsilon$-minimizer of $h$. As before, $h$ further satisfies a Hölder growth lower bound.

Lemma 3.4 The objective $h$ has $(\alpha^{p/q} \epsilon^{1-p/q}, p)$-Hölder growth on $B(x_*, D)$. Hence, $h$ is Hölder smooth and has Hölder growth with exponent $p$. Then (A3) ensures

$$T < K(h(x_0) - h(x_*), \epsilon, \alpha^{p/q} \epsilon^{1-p/q}) = K(f(f(x_0) - f_*, \epsilon, \alpha^{p/q} \epsilon^{1-p/q}).$$

3.2.1 Proof of Lemma 3.4

This proof follows the same general approach used in proving Lemma 3.3. For any $x \in B(x_*, D)$ and $\bar{x} \in \{x_0, \ldots, x_T\} \cup X_*$, $h_{\bar{x}}$ lies above our claimed growth bound: Any $x$ with $\|x - x_*\| > (\epsilon/\alpha)^{1/q}$ has

$$h_{\bar{x}}(x) \geq f(x) \geq f_* + \alpha \|x - x_*\|^q \geq f_* + \alpha^{p/q} \epsilon^{1-p/q} \|x - x_*\|^p.$$ 

Any $x$ with $\|x - x_*\| \leq (\epsilon/\alpha)^{1/q}$ and $\bar{x} = x_k$ for some $k$ has

$$h_{x_k}(x) \geq h_{x_k}(x'_k) \geq f_* + \epsilon \geq f_* + \alpha^{p/q} \epsilon^{1-p/q} \|x - x_*\|^p$$

and any $\bar{x} \in X_*$ has

$$h_{\bar{x}}(x) = f_* + \langle g(\bar{x}), x - \bar{x} \rangle + \frac{L}{\eta + 1} \|x - \bar{x}\|^{\eta + 1} \geq f_* + \frac{\epsilon}{(\epsilon/\alpha)^{p/q}} \text{dist}(x, X_*)^p.$$ 

Hence, $\min_{k \in \{0, \ldots, T, \}} \{h_k(x)\}$ satisfies our claimed Hölder growth lower bound. Since this lower bound is convex, the convex envelope $h$ must also satisfy the Hölder bound (3).

4 Conclusion

For a wide class of first-order methods, we have shown that conducting separate analyses under different growth conditions (or lack thereof) is unnecessary. Given a convergence rate assuming a growth bound, general convergence guarantees and guarantees relying on weaker growth bounds follow from our rate lifting theorems. Applying restarting schemes like [29] gives the converse, allowing improved growth
bounded convergence rates to be derived from general rates. As a result, with only minor algorithmic modifications, it suffices to analyze the first-order method in any one of these settings. Then meta-arguments can be made to derive the convergence rates in all the other cases.

The analysis machinery developed herein applies to any Hölder smooth optimization problem and allows for backward steps like those used by the proximal point method. One interesting future direction based on these results would be the analysis of composite optimization problems, minimizing the sum of two functions $f + g$. Such structured objectives are commonly addressed using proximal gradient methods or splitting algorithms. These methods are known to speed up under the same Hölder growth conditions considered here and so generalizing our rate lifting theorems may be possible.

Acknowledgements This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650441. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study. We thank the anonymous referees and associate editor for their helpful and constructive comments.

A Detailed Application to the Proximal Point Method

First, we verify (A2) holds for the proximal point method.

**Lemma A.1** For any minimizer $x_*$, (A2) holds with $D = \|x_0 - x_*\|$.

**Proof** The proximal operator prox$_{\rho, f}(\cdot)$ is nonexpansive [26]. Then since any minimizer $x_*$ of $f$ is a fixed point of prox$_{\rho, f}(\cdot)$, the distance from each iterate to $x_*$ must be nonincreasing. □

Assuming sharpness (Hölder growth with $p = 1$) facilitates a finite termination bound on the number of iterations before an exact minimizer is found (see [7] for a more general proof and discussion of this finite result). Below we compute the resulting function $K(f(x_0) - f(x_*), \alpha, \epsilon)$ that satisfies (A3).

**Lemma A.2** Consider any convex $f$ satisfying (3) with $p = 1$. Then for any $\epsilon > 0$, the proximal point method with step size $\rho > 0$ will find an $\epsilon$-minimizer $x_k$ with

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{f(x_0) - f(x_*) - \epsilon}{\rho \alpha^2}.$$ 

**Proof** The optimality condition of the proximal subproblem is $(x_{k+1} - x_k)/\rho \in \partial f(x_{k+1})$. Hence, convexity ensures

$$\|x_{k+1} - x_k\|/\|x_{k+1} - x_*\|/\rho \geq ((x_{k+1} - x_k)/\rho, x_{k+1} - x_*) \geq f(x_{k+1}) - f(x_*).$$

Then supposing that $x_{k+1} \neq x_*$, the sharp growth bound ensures

$$\|x_{k+1} - x_k\| \geq \rho (f(x_{k+1}) - f(x_*))/\|x_{k+1} - x_*\| \geq \rho \alpha.$$
Noting the proximal subproblem is $\rho$-strongly convex, the objective has
\[
f(x_{k+1}) \leq f(x_k) - \frac{\|x_{k+1} - x_k\|^2}{\rho} \leq f(x_k) - \rho \alpha^2,
\]
giving a constant decrease at each iteration until a minimizer is found. □

Applying Corollary 2.1 and 2.2, we recover every other setting’s convergence rates from the $p = 1$ rate. For $q > 1$-growth, we reach $\epsilon$-accuracy after
\[
\sum_{n=0}^{N-1} K(2^n+1\epsilon, 2^n\epsilon, \alpha^{1/q}(2^n\epsilon)^{1-1/q}) = \sum_{n=0}^{N-1} \frac{(2^n\epsilon)^{2/q-1}}{\rho \alpha^{2/q}}.
\]
This recovers the three major known regimes of the proximal point methods behavior. For $1 < q < 2$, this is converging superlinearly, for $q = 2$ converging linearly, and for $q > 2$, converging sublinearly.

### B Detailed Application to the Polyak Subgradient Method

Much like the proximal point method, the distance from each iterate of the subgradient method to a minimizer is nonincreasing when a sufficiently careful step size is employed. Below we consider the Polyak step size.

**Lemma B.1** For any minimizer $x_*$, (A2) holds with $D = \|x_0 - x_*\|$.

**Proof** The convergence of the projected subgradient method is governed by the following
\[
\|x_{k+1} - x_*\|^2 \leq \|x_k - \rho_k g_k - x_*\|^2
\]
\[
= \|x_k - x_*\|^2 - 2\langle \rho_k g_k, x_k - x_* \rangle + \rho_k^2 \|g_k\|^2
\]
\[
\leq \|x_k - x_*\|^2 - 2\rho_k (f(x_k) - f(x_*)) + \rho_k^2 \|g_k\|^2
\]
\[
= \|x_k - x_*\|^2 - \frac{(f(x_k) - f(x_*))^2}{\|g_k\|^2}
\]
\[
\leq \|x_k - x_*\|^2 - \frac{(f(x_k) - f(x_*))^2}{L^2}, \tag{10}
\]
where the first inequality uses the convexity of $f$ and the second uses our assumed subgradient bound. Thus, the distance from each iterate to $x_*$ is nonincreasing. □

As an alternative step size example, consider the step size $\rho_k = 2\epsilon/\|g_k\|^2$. Then the distance to optimal strictly decreases every step until an $\epsilon$-minimizer is found as
\[
\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\langle \rho_k g_k, x_k - x_* \rangle + \rho_k^2 \|g_k\|^2
\]
\[
\leq \|x_k - x_*\|^2 - 2\epsilon \frac{f(x_k) - f(x_*) - \epsilon}{L^2}.
\]
Below we give a simple proof that the projected subgradient method under \( L \)-Lipschitz continuity and quadratic growth finds an \( \epsilon \)-minimizer within \( O(L^2/\alpha \epsilon) \) iterations.

**Lemma B.2** Consider any convex \( f \) satisfying (3) with \( p = 2 \). Then for any \( \epsilon > 0 \), the subgradient method with the Polyak step size will find an \( \epsilon \)-minimizer \( x_k \) with
\[
k \leq K(f(x_0) - f(x_\star), \epsilon, \alpha) = \frac{8L^2}{\alpha \epsilon}.
\]

**Proof** This convergence rate follows by noting (10) implies the objective gap will halve \( f(x_k) - f(x_\star) \leq (f(x_0) - f(x_\star))/2 \) after at most
\[
\frac{4L^2 \|x_0 - x_\star\|^2}{(f(x_0) - f(x_\star))^2} \leq \frac{4L^2}{\alpha(f(x_0) - f(x_\star))}
\]
iterations. Iterating this argument, a \( 2^{-N} (f(x_0) - f(x_\star)) \)-minimizer is found with
\[
k \leq \sum_{n=0}^{N-1} \frac{4L^2}{2^{-n} \alpha(f(x_0) - f(x_\star))} \leq \frac{8L^2}{2^{-N} \alpha(f(x_0) - f(x_\star))}.
\]

Considering \( N = \lceil \log_2((f(x_0) - f(x_\star))/\epsilon) \rceil \), gives our claimed bound on the number of iterations needed to find an \( \epsilon \)-minimizer. \( \Box \)

Rather than relying on quadratic growth, assuming sharpness (Hölder growth with \( p = 1 \)) allows us to derive a linear convergence guarantee. Below we compute the resulting function \( K(f(x_0) - f(x_\star), \alpha, \epsilon) \) that satisfies (A3).

**Lemma B.3** Consider any convex \( f \) satisfying (3) with \( p = 1 \). Then for any \( \epsilon > 0 \), the subgradient method with the Polyak step size will find an \( \epsilon \)-minimizer \( x_k \) with
\[
k \leq K(f(x_0) - f(x_\star), \epsilon, \alpha) = \frac{4L^2}{\alpha^2} \log_2 \left( \frac{f(x_0) - f(x_\star)}{\epsilon} \right).
\]

**Proof** Again, this convergence rate follows by noting (10) implies the objective gap will halve \( f(x_k) - f(x_\star) \leq (f(x_0) - f(x_\star))/2 \) after at most
\[
\frac{4L^2 \|x_0 - x_\star\|^2}{(f(x_0) - f(x_\star))^2} \leq \frac{4L^2}{\alpha^2}
\]
iterations, which immediately establishes the claimed rate. \( \Box \)

**C Application to Dual Averaging**

Dual averaging methods following in the line of the primal–dual methods proposed by Nesterov [24] offer to improve on gradient methods like (5) and (6) by using an aggregate of the first-order models seen. Chen et al. [4] propose an optimal regularized
dual averaging method (and a multistage variant) attaining the optimal rate for both $L$-smooth and nonsmooth $M$-Lipschitz problems with or without a quadratic growth bound. Formally they assume strong convexity rather than quadratic growth, but the proof of Theorem 3 therein only relies on quadratic growth (invoked shortly after equation (38)). Their theorem shows optimal convergence under $\alpha$-quadratic growth at a rate of

$$K(\Delta, \epsilon, \alpha) = \frac{f}{\alpha} \log(\Delta / \epsilon) + \frac{1024M^2}{\alpha \epsilon}.$$ 

Our theory shows their dual averaging method is in fact optimal under the wider setting of Hölder growth. Applying our Corollary 2.2, we see that their dual averaging scheme (which already restarts once the objective gap is halved) is optimal both for smooth and nonsmooth problems under Hölder growth: When $L > 0$ but $M = 0$

$$N-1 \sum_{n=0}^{N-1} K(2^{n+1} \epsilon, 2^n \epsilon, \alpha^{2/q} (2^n \epsilon)^{1-2/q}) = \sum_{n=0}^{N-1} 4 \sqrt{\frac{L}{\alpha^{2/q} (2^n \epsilon)^{1-2/q}}} \log(2) = O \left( \frac{L^{1/2}}{\alpha^{1/q} \epsilon^{1/2-1/q}} \right)$$

and when $L = 0$ and $M > 0$

$$N-1 \sum_{n=0}^{N-1} K(2^{n+1} \epsilon, 2^n \epsilon, \alpha^{2/q} (2^n \epsilon)^{1-2/q}) = \sum_{n=0}^{N-1} \frac{1024M^2}{\alpha^{2/q} (2^n \epsilon)^{2-2/q}} = O \left( \frac{M^2}{\alpha^{2/q} \epsilon^{2-2/q}} \right).$$

**D Application to the Proximal Bundle Method**

Bundle methods have a long history throughout the nonsmooth optimization literature. These methods build sufficiently accurate models of an objective to ensure descent, despite its nonsmoothness. Until recently, convergence rate theory was lacking in terms of the total number of subgradient evaluations required. Kiwiel [11] showed a $O(1/\epsilon^3)$ convergence rate in 2000, and then, in 2017, Du and Ruszczyński [6] showed a $O(1/\epsilon \alpha^2)$ rate under quadratic growth, up to logarithmic terms. Both rates fall short of the optimal $O(1/\epsilon^2)$ and $O(1/\epsilon \alpha)$.

The boundness condition (A2) is verified by [32, (7.64)]. Then applying our theory shows these rates are similarly suboptimal as Theorem 2.1 shows Du and Ruszczyński’s $O(1/\epsilon \alpha^2)$ rate implies Kiwiel’s $O(1/\epsilon^3)$ rate and a restarting argument shows the converse. Applying our Theorem 2.2 gives new rates for the bundle method in every $2 < p < \infty$ growth setting of $O(1/\alpha^{2/p} \epsilon^{3-2/p})$. Subsequent to this work, in [5], the author directly derived these bounds (as well as bounds using alternative step size rules, fixing their suboptimality). Note direct proofs of these rates ended up being nontrivial.
E Application to the Frank–Wolfe Method

Frank–Wolfe [8] repeatedly solves linear optimization subproblems (based on gradient evaluations) over a given compact constraint set $Q$. A bound $D$ on the diameter of this compact constraint set suffices for (A2). Recently, a variant of Frank–Wolfe including (fractional) away-steps and periodic restarting was analyzed under various Hölder smoothness conditions and any $p \geq 2$-Hölder growth bound in [9, Theorem 6.2]. They assume the constraint set $Q$ satisfies a certain $\delta$-scaling property (Definition 3.3 therein, which is satisfied, for example, by all polytopes) and utilize curvature bounds (away and standard) on $f$, which are both bounded by the smoothness constant $L$ when $\eta = 1$.

In terms of the notation and numbering of [9], their equation (15) is missing a power of $r$ in the $\mu$ term in both of the last two inequalities (see Lemma 3.6, which is being invoked there and requires $\mu = c/\delta$ to be raised to the power of $r = 1/(1 - \theta)$). This missing exponent can be carried forward through their arguments directly, only modifying their final convergence rates by including a power of $r$ in $\mu$. Converting their quantities to match the notations of this paper, the corrected convergence rates from their Theorem 5.1 are $O\left(\frac{1}{\alpha^{1/p} \epsilon^{1/2}}\right)$ when $p > 2$ and $O\left(\log(1/\epsilon)/\alpha\right)$ when $p = 2$. Applying our Corollary 2.2 with their quadratic growth guarantee to the $p > 2$ case proves an improved dependence on $\epsilon$ of $O\left(\frac{1}{\alpha^{2/p} \epsilon^{1/2}}\right)$ when $p = 2$.

F Application to the Universal Fast Method with Restarting

For any $(L, \eta)$-Hölder smooth, convex optimization problem, Nesterov’s universal fast gradient method [25] attains the optimal convergence rate and the restarting scheme of [29] improved this convergence theory under any $(\alpha, p)$-Hölder growth. Let us briefly verify the claims in Table 1 for any $L$-smooth function. Nesterov’s Theorem 3 shows a rate of $O\left(\sqrt{L D^2} / \epsilon\right)$ for any bound $D > 0$ on the iterates which Renegar and Grimm’s Corollary 10 shows (via restarting) implies faster rates of $O\left(\sqrt{L/\alpha} \log(\Delta/\epsilon)\right)$ if $p = 2$ and $O\left(\sqrt{L/\alpha^{1/p} \epsilon^{1/2-1/p}}\right)$ otherwise. Applying our theory provides the reverse implication: the $p = 2$ setting implies rates for $p > 2$ and the general setting.

References

1. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17(4), 1205–1223 (2007). https://doi.org/10.1137/050644641
2. Bolte, J., Nguyen, T.P., Peyrouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. 165(2), 471–507 (2017). https://doi.org/10.1007/s10107-016-1091-6
3. Burke, J.V., Ferris, M.C.: Weak sharp minima in mathematical programming. SIAM J. Control. Optim. 31(5), 1340–1359 (1993). https://doi.org/10.1137/0331063
4. Chen, X., Lin, Q., Peña, J.: Optimal regularized dual averaging methods for stochastic optimization. In: Advances in Neural Information Processing Systems (2012)
5. Díaz, M., Grimmer, B.: Optimal convergence rates for the proximal bundle method. SIAM J. Optim. (to appear) (2021)
6. Du, Y., Ruszczyński, A.: Rate of Convergence of the Bundle Method. J. Optim. Theory Appl. 173(3), 908–922 (2017). https://doi.org/10.1007/s10957-017-1108-1
7. Ferris, M.: Finite termination of the proximal point algorithm. Math. Program. 50, 359–366 (1991). https://doi.org/10.1007/BF01594944
8. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist. Q. 3(1–2), 95–110 (1956). https://doi.org/10.1002/nav.3800030109
9. Kerdreux, T., d’Aspremont, A., Pokutta, S.: Restarting Frank–Wolfe: faster rates under Hölderian error bounds. J. Optim. Theory Appl. 192(3), 799–829 (2022). https://doi.org/10.1007/s10957-021-01989-7
10. Kiwiel, K.C.: Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities. Math. Program. 69(1–3), 89–109 (1995). https://doi.org/10.1007/BF01585554
11. Kiwiel, K.C.: Efficiency of proximal bundle methods. J. Optim. Theory Appl. 104(3), 589–603 (2000). https://doi.org/10.1023/A:1004689609425
12. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l’institut Fourier 48(3), 769–783 (1998)
13. Lan, G.: Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. Math. Program. 149(1), 1–45 (2015). https://doi.org/10.1007/s10107-013-0737-x
14. Lemaréchal, C.: An extension of Davidon methods to nondifferentiable problems. In: Nondifferentiable Optimization, pp. 95–109. Springer, Berlin (1975). https://doi.org/10.1007/BFb0120700
15. Lemaréchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Math. Program. 69(1–3), 111–147 (1995). https://doi.org/10.1007/BF01585555
16. Li, G., Mordukhovich, B.S., Pham, T.S.: New fractional error bounds for polynomial systems with applications to Hölderian stability in optimization and spectral theory of tensors. Math. Program. 153(2), 333–362 (2015). https://doi.org/10.1007/s10107-014-0806-9
17. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. 18(5), 1199–1232 (2018). https://doi.org/10.1007/s10208-017-9366-8
18. Liu, M., Yang, T.: Adaptive accelerated gradient converging method under Hölderian error bound condition. In: Advances in Neural Information Processing Systems 30, pp. 3104–3114. Curran Associates, Inc. (2017)
19. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles 117, 87–89 (1963)
20. Łojasiewicz, S.: Sur la géométrie semi-eti sous-analytique. Annales de l’institut Fourier 43(5), 1575–1595 (1993)
21. Necula, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program. 175(1), 69–107 (2019). https://doi.org/10.1007/s10107-018-1232-1
22. Nemirovskii, A., Nesterov, Y.: Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. 25(3–4), 21–30 (1986)
23. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, 1st edn. Springer Publishing Company, Incorporated (2004)
24. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. 120(1), 221–259 (2009). https://doi.org/10.1007/s10107-007-0149-x
25. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. 152, 381–404 (2015). https://doi.org/10.1007/s10107-014-0790-0
26. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. 1(3), 127–239 (2014). https://doi.org/10.1561/2200000003
27. Polyak, B.T.: Minimization of unsmooth functionals. USSR Comput. Math. Math. Phys. 9(3), 14–29 (1969). https://doi.org/10.1016/0041-5553(69)90061-5
28. Polyak, B.T.: Sharp minima. Institute of Control Sciences Lecture Notes, Moscow, USSR, Presented at the IIASA Workshop on Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria. (1979)
29. Renegar, J., Grimmer, B.: A simple nearly optimal restart scheme for speeding up first-order methods. Found. Comput. Math. 22(1), 211–256 (2022). https://doi.org/10.1007/s10208-021-09502-2
30. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. 14(5), 877–898 (1976). https://doi.org/10.1137/0314056
31. Roulet, V., d’Aspremont, A.: Sharpness, restart, and acceleration. SIAM J. Optim. 30(1), 262–289 (2020). https://doi.org/10.1137/18M1224568
32. Ruszczynski, A.: Nonlinear Optimization. Princeton University Press, Princeton (2006)
33. Wolfe, P.: A method of conjugate subgradients for minimizing nondifferentiable functions. In: Nondifferentiable Optimization, pp. 145–173. Springer, Berlin (1975). https://doi.org/10.1007/BFb0120703
34. Yang, T., Lin, Q.: RSG: Beating subgradient method without smoothness and strong convexity. J. Mach. Learn. Res. 19(6), 1–33 (2018)

Publisher's Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.