Inexactly Smooth Performance Estimation and New Optimized Gradient Methods

Aaron Zoll* Benjamin Grimmer†

Abstract

We consider a general class of "inexactly smooth" convex functions, providing a universal model capturing as special cases L-smooth, M-Lipschitz, and Hölder smooth functions, and any combination thereof. Such functions possess a calculus closely following that of smooth functions. Our main results provide inexactly smooth functions with interpolation theorems that are necessary and sufficient up to modest universal constants. These enable analysis of first-order methods for any inexactly smooth convex problem class via solving convex Performance Estimation Problems (PEPs). Further, these enable the extension of Drori and Taylor [9]'s constructive approach to algorithm design. From this, we derive an exactly minimax optimal method for $(\beta, 0)$-Hölder smooth problems, methods with the best-known convergence guarantees up to constants for any (β, p)-Hölder smooth convex minimization, and a new universal fast backtracking method for any inexactly smooth convex problem.

1 Introduction

First-order methods have found success in a wide range of convex optimization tasks, ranging from smooth minimization to nonsmooth minimization. First-order methods are particularly effective at modern, large-scale unconstrained optimization problems of the form

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)$$

where computations of (sub)gradients remain tractable while other oracle models become expensive. Since the 1970s, algorithms tailored to the structure of the given optimization problem class (smooth, nonsmooth, etc.) have been developed with highly optimized performance, often being minimax optimal. As classic examples, see [24, 25].

Alas, these optimized methods are typically only guaranteed to be performant on their target class of problems and arise from analysis techniques tailored to that class. For example, methods may be specialized to smooth convex problems or Lipschitz convex problems. Here, our goal is to develop analytic tools, in particular "Performance Estimation Problems," uniformly applicable to algorithm design over a wide range of problem classes.

To this end, we are motivated by the universal methods and analysis of [6] and [26]. Therein, they design first-order methods given access to a first-order oracle returning $(f(x), g)$ with $g \in \partial f(x)$ for any query x. Here $\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x\rangle \ \forall y\}$ denotes the convex subdifferential. In particular, they considered any convex problem with Hölder continuous subgradients, which we refer to here as (β, p)-Hölder smoothness,

$$\|g_x - g_y\| \leq \beta \|x - y\|^p \qquad \forall x, y \in \mathbb{R}^d,\ g_x \in \partial f(x),\ g_y \in \partial f(y) \tag{1.1}$$

*Johns Hopkins University, Department of Applied Mathematics and Statistics, azoll1@jhu.edu
†Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

for given $\beta > 0$ and $p \in [0, 1]$. Nesterov's Universal Fast Gradient Method (UFGM) [26] ensures

$$f(x_N) - f_\star \leq \frac{2^{1+2p}\beta\|x_0 - x_\star\|^{1+p}}{N^{\frac{1+3p}{2}}} \tag{1.2}$$

universally, which is within a constant factor of the minimax optimal rate [23] for each $p \in [0, 1]$. When $p = 1$, Hölder smoothness corresponds to the classical smoothness assumption of $\beta = L$-Lipschitz gradient. When $p = 0$, this class corresponds to functions with subgradients differing by at most β. This is closely related, but distinct, from the classical nonsmooth assumption of having $M = \beta$-Lipschitz function values. Considering $p \in (0, 1)$ interpolates between these settings.

The key analysis tool leveraged by [6] and [26] is that Hölder smoothness (1.1) implies the following inexact quadratic upper bound holds for all $x, y \in \mathbb{R}^d$, $g_x \in \partial f(x)$ and any $\delta \geq 0$

$$f(y) \leq f(x) + \langle g_x, y - x\rangle + \frac{\left(\frac{1-p}{1+p} \cdot \frac{1}{2\delta}\right)^{\frac{1-p}{1+p}} \beta^{\frac{2}{1+p}}}{2}\|y - x\|^2 + \delta. \tag{1.3}$$

This result allows one to approach any Hölder smooth setting with similar methods and proof techniques to the smooth setting ($p = 1$) where this bound holds without the additive error δ. Hence, such inexact bounds are an established means to universalizing the first-order method toolbox.

Generalizing this key analysis tool, here we propose a family of "inexactly smooth" functions. We say a convex function is $L(\cdot)$-inexactly smooth for some function $L\colon [0, \infty) \to (0, \infty]$, if for any $x, y \in \mathbb{R}^d$, $g_x \in \partial f(x)$, and $\delta \geq 0$,

$$f(y) \leq f(x) + \langle g_x, y - x\rangle + \frac{L(\delta)}{2}\|y - x\|^2 + \delta. \tag{1.4}$$

If $L(\delta) = \left(\frac{1-p}{1+p} \cdot \frac{1}{2\delta}\right)^{\frac{1-p}{1+p}} \beta^{\frac{2}{1+p}}$, this family contains all (β, p)-Hölder smooth functions. This can additionally model functions generated by sums of smooth and nonsmooth components, which have also received notable theoretical interest recently [7, 15, 17, 31]. Given a sum $f = f_0 + f_1$, with f_0 having subgradients differ by at most β_0 and f_1 having β_1-Lipschitz gradient, inexact smoothness holds with $L(\delta) = \beta_1 + \beta_0^2/(2\delta)$. In Section 3, we discuss more properties and examples within this inexactly smooth model of functions.

In this paper, we develop a principled approach to the design and analysis of algorithms for inexactly smooth convex minimization. The foundational works of [10] and [29, 30] introduced Performance Estimation Problems (PEPs) as a principled way to design and analyze first-order methods over traditional problem classes. See PEPit [14] and the many examples therein. PEPs are mathematical programs (often semidefinite programs) that compute a worst-case problem instance from a given class (e.g., L-smooth convex problems) for a given algorithm. Dually, PEPs provide a best possible convergence proof for a given algorithm using a given set of inequalities. Agreement of these primal and dual perspectives relies on related "Interpolation Theorems."

Our Contributions. This work has two primary objectives. First, we provide foundations (i.e., inexactly smooth calculus, interpolation theorems) for applying modern algorithm design tools to families of inexactly smooth functions. Second, we utilize these to provide PEPs and apply the constructive approach of [9] to algorithm design for such families. Namely, we show

- **Calculus of Inexactly Smooth Functions** In Section 2, we derive characterizations of inexactly smooth functions and their respective conjugates, as well as associated calculus rules.

- **Interpolation Theory** Section 3 begins by presenting necessary and sufficient conditions for observations to be interpolable by an inexactly smooth function. Theorem 3.3 gives our most general interpolation theory, which is tight up to a small universal constant. Theorem 3.5 establishes that these conditions are exact for $(\beta, 0)$-Hölder smooth functions.
- **Inexactly Smooth Performance Estimation** We conclude Section 3 by using our interpolation theory to define a convex formulation for analyzing the worst-case performance of an algorithm. Note, unlike most existing PEP work, these convex programs may not be semidefinite programs. Regardless, Theorem 3.6 establishes a strong duality result.
- **Algorithm Design** Section 4 leverages our interpolation theory to construct optimized algorithms. Algorithm 1 is minimax optimal among convex, $(\beta, 0)$-Hölder smooth functions, Algorithm 2 is an asymptotically optimized gradient method for general Hölder smooth, convex minimization, and Algorithm 3 is a universal method for any inexactly smooth convex problem requiring only input of a target accuracy ε and an iteration budget N. The associated convergence guarantees are given in Theorems 4.2, 4.3, and 4.5, respectively.

Note the inexactness considered here (1.4) is inexactness in attainment of smoothness-type quadratic upper bounds. This is distinct from inexactness due to gradient calculations or noise, which have separately been studied in the PEP literature [4, 5, 13, 22].

2 Preliminaries and Inexactly Smooth Function Calculus

We begin with preliminaries and then discuss Hölder smooth functions as our motivating instance of inexactly smooth functions in Section 2.1. Then Section 2.2 develops a general calculus for $L(\cdot)$-inexactly smooth functions which may be of independent interest.

Throughout, we consider closed, convex, proper functions $f : \mathbb{R}^d \to (-\infty, \infty]$. Let $\langle \cdot, \cdot \rangle$ denote the standard Euclidean inner product with the associated two-norm $\|\cdot\|$. We denote the domain of f by $\mathrm{dom}(f)$ and its subdifferential as previously introduced by $\partial f(x)$. We denote $\mathrm{epi}(f) := \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq t\}$ as the epigraph of f, for which the function f is convex if and only if the set $\mathrm{epi}(f)$ is convex. Similarly, we denote $\mathrm{hypo}(f) := \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : f(x) \geq t\}$ as the hypograph. The Fenchel conjugate of f is $f^*(g) = \sup_{x \in \mathbb{R}^d}\{\langle g, x\rangle - f(x)\}$, which satisfies the Fenchel-Young inequality, which holds with equality exactly when $g \in \partial f(x)$,

$$f(x) + f^*(g) \geq \langle g, x\rangle \qquad \forall x, g \in \mathbb{R}^d. \tag{2.1}$$

2.1 Hölder Smooth Functions

A core motivation for our work on general inexactly smooth functions is the special case of (β, p)-Hölder smooth functions (i.e., having Hölder continuous (sub)gradient as defined in (1.1)). The extreme cases of $p = 1$ and $p = 0$ are particularly fundamental. When $p = 1$, this is exactly f having β-Lipschitz gradient and corresponds to $L(\delta) = \beta$-inexact smoothness. When $p = 0$, this bounds the difference between two subgradients by β and corresponds to $L(\delta) = \frac{\beta^2}{2\delta}$.

Devolder, Glineur, and Nesterov [6] connected Hölder smoothness for any $p \in [0, 1]$ to $L(\delta)$-inexact smoothness. Below we formalize this relationship, where the coefficient $\left(\frac{p+1}{2p}\right)^p$ is defined at $p = 0$ by its limiting value of 1 as $p \to 0$. By convention, we take $1/\infty = 0$ and $1/0 = \infty$.

Proposition 2.1. *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be closed convex and proper, and $L(\delta) = \left(\frac{1-p}{1+p}\frac{1}{2\delta}\right)^{\frac{1-p}{1+p}} \beta^{\frac{2}{1+p}}$. Then the conditions*

1. $\|g_x - g_y\| \leq \beta \|x-y\|^p, \quad \forall\ x, y \in \mathrm{dom}(f)$ *and* $g_x \in \partial f(x), g_y \in \partial f(y)$,
2. $f(y) \leq f(x) + \langle g_x, y-x\rangle + \frac{L(\delta)}{2}\|x-y\|^2 + \delta, \quad \forall\ x, y \in \mathrm{dom}(f)$, $g_x \in \partial f(x)$, *and* $\delta \geq 0$,
3. $f(y) \geq f(x) + \langle g_x, y-x\rangle + \frac{1}{2L(\delta)}\|g_x - g_y\|^2 - \delta, \quad \forall\ x, y \in \mathrm{dom}(f)$, $g_x \in \partial f(x), g_y \in \partial f(y)$, *and* $\delta \geq 0$,
4. $\|g_x - g_y\| \leq \beta \left(\frac{p+1}{2p}\right)^p \|x-y\|^p, \quad \forall\ x, y \in \mathrm{dom}(f)$ *and* $g_x \in \partial f(x), g_y \in \partial f(y)$.

are related by the implications 1. $\implies$ 2. $\iff$ 3. $\implies$ 4.

Proof. The proof of [21, Lemma 1] proves (1. $\implies$ 2. $\implies$ 3. $\implies$ 4.) for all $p \in (0,1)$. Considering limits as $p \to 0$ and $p \to 1$ extends this to $p \in [0,1]$. In Lemma 2.4, we will provide the final needed implication, establishing (2 $\iff$ 3.). □

Note that the above implications are equivalent up to this factor of $\left(\frac{p+1}{2p}\right)^p \leq 1.263$. This factor is tight. Hence, a small constant factor gap is fundamental when approximating the family of Hölder smooth functions via smooth quadratic bounds with additive errors for $p \in (0,1)$. Tightness is demonstrated by the following function that satisfies conditions 2. through 4., but holds with equality in 4. when $x = 0$ and $y = 1$,

$$f_{\beta,p}(x) := \begin{cases} 0 & x < 0 \\ \frac{1}{2}\beta\left(\frac{p+1}{2p}\right)^p x^2 & 0 \leq x \leq 1 \\ \beta\left(\frac{p+1}{2p}\right)^p \left(x - \frac{1}{2}\right) & x > 1. \end{cases} \tag{2.2}$$

2.2 Calculus of $L(\cdot)$-inexactly smooth functions

Here, we develop a calculus for $L(\cdot)$-inexactly smooth functions previously defined in (1.4). Complementing this, we say a convex function f is $\mu(\cdot)$-inexactly strongly convex if

$$f(y) \geq f(x) + \langle g_x, y-x\rangle + \frac{\mu(\delta)}{2}\|x-y\|^2 - \delta, \quad \forall x, y \in \mathbb{R}^d,\ g_x \in \partial f(x). \tag{2.3}$$

Our development of characterizations of these inexact quantities closely follows the classical developments of smoothness and strong convexity. Consequently, we defer proofs to Appendix A when they are analogous. Throughout, we assume the following structure on $L(\cdot)$.

Assumptions on $L(\delta)$: Inexact smoothness functions $L\colon [0,\infty) \to (0,\infty]$ are assumed finite for $\delta > 0$, convex, lower semicontinuous, nonincreasing, and such that $-1/L(\cdot)$ is convex.

We note the following equivalent convexity characterization.

Lemma 2.2. *For any $L\colon [0,\infty) \to (0,\infty]$ that is convex, lower semicontinuous, and nonincreasing, one has that*

$$-1/L(\delta) \text{ is convex} \quad \iff \quad sL^{\leftarrow}(s) \text{ is convex},$$

where we define the left inverse $L^{\leftarrow}(s) := \inf\{\delta \geq 0 : L(\delta) \leq s\}$.

Proof. Consider the set $\mathcal{S} := \left\{\left(\frac{1}{\alpha}, \frac{\delta}{\alpha}\right) : (\delta, \alpha) \in \text{hypo}(1/L),\ \alpha > 0\right\}$. We claim that $\mathcal{S}$ is the epigraph of $s \mapsto sL^{\leftarrow}(s)$. The condition $(\delta, \alpha) \in \text{hypo}(1/L)$ means $\alpha \leq 1/L(\delta)$, equivalently $L(\delta) \leq 1/\alpha$. Considering the definition of $\mathcal{S}$ and writing $s := 1/\alpha$ and $z := \delta/\alpha = \delta s$, the previous inequality states there exists $\delta \geq 0$ such that $L(\delta) \leq s$ and $z = \delta s$. Since $L(\cdot)$ is nonincreasing and $1/L(\cdot)$ is nondecreasing, we may rewrite

$$\mathcal{S} = \{(s, z) :\ \exists \delta > 0 \text{ s.t. } L(\delta) \leq s \text{ and } z \geq \delta s\}.$$

Fix some $s > 0$. Among all $\delta \geq 0$ with $L(\delta) \leq s$, the smallest possible value of δs is

$$s \cdot \inf\{\delta \geq 0 :\ L(\delta) \leq s\} = s\, L^{\leftarrow}(s).$$

Hence $\mathcal{S}$ is the epigraph of the function $s \mapsto sL^{\leftarrow}(s)$.

The following equivalences hold between convexity of functions and convexity of their associated epigraphs/hypographs

$$-1/L(\delta) \Longleftrightarrow \text{hypo}(1/L(\delta)) \Longleftrightarrow \text{epi}(sL^{\leftarrow}(s)) \Longleftrightarrow sL^{\leftarrow}(s),$$

where the first and last implications note convexity and concavity are equivalent to convexity of epigraphs and hypographs, and the middle implication holds as $\mathcal{S}$ is a perspective transformation of the hypograph intersected with a halfplane [2, Section 2.3.3]. □

First, we note that one can always relax an inexact smoothness function $L(\cdot)$ (under these assumptions) into the sum of two inexact smoothness functions previously seen for $p = 1$ and $p = 0$ Hölder smoothness, $L(\delta) \leq \beta_1 + \beta_0^2/(2\delta)$. Therefore, these extremal cases of Hölder smoothness are also extremal among inexactly smooth functions.

Lemma 2.3. *If $L(\cdot)$ satisfies our assumptions, then for any $a > 0$, $L(\delta) \leq \frac{aL(a)}{\delta} + L\,(a)$.*

Proof. By concavity of $1/L(\cdot)$ it holds that for any $a > 0$ and $t \in [0, 1]$ that

$$\frac{1}{L(ta)} = \frac{1}{L(ta + (1-t)0)} \geq \frac{t}{L(a)} + \frac{1-t}{\lim_{\delta \to 0^+} L(\delta)} \geq \frac{t}{L(a)},$$

where the last inequality comes from the positivity of $L(\cdot)$. Letting $\delta = ta$, this implies that $L(\delta) \leq \frac{aL(a)}{\delta}$ for any $\delta \in [0, a]$. Then by monotonicity of $L(\cdot)$, $L(\delta) \leq L(a)$ holds for all $\delta \geq a$. Summing the nonnegative upper bounds from these two regimes gives the claim. □

Next, we derive useful equivalent characterizations of $L(\cdot)$-inexact smoothness. The following lemma shows that among convex functions f, inexact smoothness can equally be viewed as satisfying a cocoercivity-type condition or as the conjugate f^* being inexactly strongly convex with $\mu(\cdot) = 1/L(\cdot)$. These follow similarly to the standard textbook fact for (non-inexactly) smooth convex functions.

Lemma 2.4. *Let f be closed, convex, and proper. The following are equivalent:*

1. $f(y) \geq f(x) + \langle g_x, y - x\rangle + \frac{1}{2L(\delta)}\|g_y - g_x\|^2 - \delta, \quad \forall \delta \geq 0,\ \forall x, y$ *with* $g_x \in \partial f(x),\ g_y \in \partial f(y)$,
2. *f is $L(\cdot)$-inexactly smooth,*
3. *f^* is $1/L(\cdot)$-inexactly strongly convex.*

The following three lemmas provide collections of calculus rules for constructing inexactly smooth functions and quantifying their inexact smoothness $L(\cdot)$. As with the above result, their proofs are analogous to the standard smooth and convex facts and hence, deferred to Appendix A. Note, by considering conjugate functions and Lemma 2.4, equivalent calculus rules for inexactly strongly convex functions could be derived directly from these.

Lemma 2.5. *Consider any convex, $L(\cdot)$-inexactly smooth function f. Then,*

1. *$\alpha f(x)$ is $\alpha L(\cdot/\alpha)$-inexactly smooth for any $\alpha > 0$,*
2. *$f(Ax - b)$ is $\|A\|_{\text{op}}^2 L(\cdot)$-inexactly smooth,*
3. *Suppose $F(x) = \inf_z f(x, z)$ is closed and proper with attainment for each x, then $F(x)$ is $L(\cdot)$-inexactly smooth.*

Lemma 2.6. *Consider any closed, convex, and proper $L_i(\cdot)$-inexactly smooth functions f_i for $i = 1, \dots, m$. Then,*

1. $\displaystyle\sum_{i=1}^m f_i$ *is* $\left(\inf_{\substack{\sum_{i=1}^m \delta_i = \delta \\ \delta_i \geq 0}} \sum_{i=1}^m L_i(\delta_i)\right)$*-inexactly smooth,*
2. $(\max_i f_i^*)^*$ *is* $\max_i L_i(\delta)$*-inexactly smooth,*
3. $\left(\displaystyle\sum_{i=1}^m f_i^*\right)^*$ *is* $\left(\sup_{\substack{\sum_{i=1}^m \delta_i = \delta \\ \delta_i \geq 0}} \sum_{i=1}^m \frac{1}{L_i(\delta_i)}\right)^{-1}$ *-inexactly smooth.*

Lemma 2.7. *Consider any convex function $h : \mathbb{R} \to \mathbb{R}$ such that $h(|t|)$ is $L(\cdot)$-inexactly smooth. Then $f(x) = h(\|x\|)$ is $L(\cdot)$-inexactly smooth.*

Rather than retaining a family of inequalities for each $\delta \geq 0$, one may consider the strengthened inequality given by optimizing over all $\delta \geq 0$. To this end, we define

$$\phi_L(s, \delta) := \frac{L(\delta)}{2} s^2 + \delta, \quad \theta_L(s) := \inf_{\delta \geq 0} \phi_L(s, \delta). \tag{2.4}$$

Below, we note a few properties of the joint function $\phi_L(s, \delta)$ and its partial minimization θ_L.

Lemma 2.8. *Suppose $L(\cdot)$ satisfies our assumptions. Then $\phi_L(s, \delta)$ as defined in* (2.4) *is jointly convex in s and δ. When attained, $\{\delta_\star(s)\} := \operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta)$ is a singleton, and if $s \neq 0$, then $L(\delta_\star(s)) < \infty$.*

Proof. Consider the function $g(u, v) = u^2/v$, which is jointly convex and decreasing in $v > 0$. Then since $1/L(\delta)$ is concave, the composition $\phi_L(s, \delta) = g(s, 2/L(\delta)) + \delta$ is jointly convex.

We now show that $\operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta)$ is a singleton for all s. If $s = 0$, this is trivial. Otherwise, without loss of generality, assume $s > 0$ (since $\phi_L(\cdot, \delta)$ is even) and suppose the convex level set, $\operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta) = [\delta_1, \delta_2]$. Because $\phi_L(s, \cdot)$ is convex, it must hold that $\phi_L(s, \delta) = \theta_L(s)$ is constant on $[\delta_1, \delta_2]$. However, this implies that

$$L(\delta) = \frac{2}{s^2}(\theta_L(s) - \delta), \quad \delta \in [\delta_1, \delta_2]$$

for fixed $s > 0$ and $\theta_L(s) > 0$. Since $L(\delta) > 0$ by our assumptions, it holds that on this interval $-1/L(\cdot)$ is strictly concave, which can only be compatible with the assumption that $-1/L(\cdot)$ is convex when $\delta_1 = \delta_2$. Therefore, it must hold that $\operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta)$ is a singleton when attained. Finally, when $s \neq 0$, since $L(\cdot)$ is finite for all $\delta > 0$, it holds that

$$L(\delta_\star(s)) \leq \frac{\min_{\delta \geq 0} \phi_L(s, \delta)}{s^2/2} \leq \frac{\phi_L(s, 1)}{s^2/2} < \infty.$$

□

Lemma 2.9. *Suppose $L(\cdot)$ satisfies our assumptions. Then*

$$\theta_L^*(u) = \sup_{\delta \geq 0} \left\{ \frac{1}{2L(\delta)} u^2 - \delta \right\},$$

and both θ_L and θ_L^* are closed, convex, proper, and even. Finally, θ_L vanishes only at 0, and for all $s \neq 0$, the subdifferential is a singleton.

Proof. First, we verify that θ_L is closed, convex, and proper. Closedness and properness follow immediately from the assumed structure of $L(\cdot)$. To verify convexity, note that $\theta_L(s)$ is the partial minimization of $\phi_L(s, \delta)$, so convexity is preserved by [28, Theorem 5.3].

Closedness and convexity of θ_L^* follow as it is a conjugate. Properness of θ_L^* follows from the properness of θ_L. The claimed formula for this conjugate follows by computing

$$\begin{aligned}
\theta_L^*(u) &= \sup_s \sup_{\delta \geq 0} \left\{ us - \delta - \frac{L(\delta)}{2} s^2 \right\} \\
&= \sup_{\delta \geq 0} \sup_s \left\{ us - \delta - \frac{L(\delta)}{2} s^2 \right\} = \sup_{\delta \geq 0} \left\{ \frac{1}{2L(\delta)} u^2 - \delta \right\}.
\end{aligned}$$

Notice that both functions are even as they only depend on their input through a squared term, so $\theta_L(s) = \theta_L(-s)$ and $\theta_L^*(u) = \theta_L^*(-u)$.

As $L(\delta)$ is finite for $\delta > 0$, it holds that $\theta_L(0) = \inf_{\delta \geq 0} \delta = 0$. If $s \neq 0$, we can show that $\theta_L(s) > 0$ by providing a uniform lower bound on $\phi_L(s, \delta) = \frac{L(\delta)}{2} s^2 + \delta$ for all $\delta \geq 0$. First, for $\delta \leq 1$, by monotonicity of $L(\cdot)$, we can bound $\phi_L(s, \delta) \geq \frac{L(1)}{2} s^2 > 0$. Now suppose $\delta > 1$. By concavity and monotonicity of $1/L(\cdot)$, there exists some $a \geq 0$ such that $L(\delta) \geq \frac{1}{\frac{1}{L(1)} + a(\delta - 1)}$. We then apply the arithmetic-geometric mean inequality to bound

$$\frac{L(\delta)}{2} s^2 + \delta \geq |s| \sqrt{\frac{2L(1)\delta}{1 + aL(1)(\delta - 1)}} \geq |s| \sqrt{\min\{2L(1), 2/a\}} > 0,$$

where the first bound substitutes the bound derived from the concavity of $1/L(\cdot)$ and the second bound considers whether $aL(1) \leq 1$ or $aL(1) > 1$. Therefore,

$$\theta_L(s) = \inf_{\delta \geq 0} \phi_L(s, \delta) \geq \min \left\{ \frac{L(1)}{2} s^2, |s| \sqrt{2L(1)}, |s| \sqrt{2/a} \right\} > 0.$$

Lastly, since $\phi_L(s, \delta)$ is closed, convex, and proper, [19, Theorem 3.101] ensures that

$$\partial \theta_L(s) = \{u : (u, 0) \in \partial \phi_L(s, \delta_\star(s)),\ \delta_\star(s) \in \operatorname*{argmin}_{\delta \geq 0} \phi_L(s, \delta)\}.$$

For $s \neq 0$, Lemma 2.8 demonstrates that $\operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta) = \{\delta_\star(s)\}$ is unique and $L(\delta_\star(s)) < \infty$. Coupled with $\phi_L(\cdot, \delta_\star(s))$ being differentiable in s at $(s, \delta_\star(s))$ since $L(\delta_\star(s)) < \infty$, it holds that $\partial \theta_L(s)$ is a singleton: non-empty by convexity of θ_L, and every element of $\partial \phi_L(s, \delta_\star(s))$ has first coordinate equal to $L(\delta_\star(s))s$, as outlined above. □

3 Inexactly Smooth Performance Estimation

This section presents the main results on the interpolation theory for $L(\cdot)$-inexactly smooth functions. Using the calculus results from Section 2.2, we state our interpolation theorems with their proofs given in Section 3.1. Subsequently, Section 3.2 uses these interpolation theorems to present a tractable convex optimization problem for estimating the worst-case performance of a given (sub)gradient method over a given family of inexactly smooth problems.

Definition 3.1. *Consider a set of observations $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{I}}$ with $x_i, g_i \in \mathbb{R}^d$ and $f_i \in \mathbb{R}$ for some index set $\mathcal{I}$. We say $\mathcal{H}$ is $L(\cdot)$-interpolable if there exists a convex, $L(\cdot)$-inexactly smooth function f such that $f(x_i) = f_i$ and $g_i \in \partial f(x_i)$ for all $i \in \mathcal{I}$.*

Our main goal is to provide necessary and sufficient conditions for interpolability. Our conditions below provide this, up to a small change in the function $L(\cdot)$ by an absolute constant. Let the "self-smoothness constant" c_L of θ_L (previously defined in (2.4)) be the smallest positive number for which θ_L is $c_L L(\cdot/c_L)$-inexactly smooth.

$$c_L = \inf\{c > 0 : \theta_L \text{ is } cL(\cdot/c)\text{-inexactly smooth}\} \tag{3.1}$$

We can write this constant as

$$c_L = \inf\{c > 0 : \forall r, s, g_s \in \partial\theta_L(s),\, \theta_L(r) \le \theta_L(s) + g_s(r-s) + c\theta_L(r-s)\}$$

since $c\theta_L(s) = c\inf_{\delta\ge 0}\left\{\frac{L(\delta)}{2}s^2 + \delta\right\} = \inf_{\delta\ge 0}\left\{\frac{cL(\delta/c)}{2}s^2 + \delta\right\}$. The following theorem notes that this self-smoothness constant is well-defined and at most two for any $L(\cdot)$ that satisfies our assumptions.

Theorem 3.2. *If $L(\cdot)$ satisfies our assumptions, then $1 \le c_L \le 2$.*

Necessary conditions for interpolation follow from any necessary condition for inexactly smooth convex functions, specialized to the points x_i. For example, the inequality (1.4) with $y = x_i$ and $x = x_j$ and the subgradient inequality $f_i \ge f_j + \langle g_j, x_i - x_j\rangle$ are necessary. In the setting of L-smooth convex functions, the interpolation theorem [30, Corollary 1] showed that the cocoercive inequality $f_i \ge f_j + \langle g_j, x_i - x_j\rangle + \frac{1}{2L}\|g_i - g_j\|^2$ for all $i, j \in \mathcal{I}$ is necessary and sufficient. The following theorem shows that the generalization of this cocoercive-type condition in Lemma 2.4 remains necessary and sufficient for inexactly smooth interpolation, up to the self-smoothness constant c_L.

Theorem 3.3. *Let $L(\cdot)$ satisfy our assumptions. If a set of observations $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{I}}$ is $L(\cdot)$-interpolable, then for each $i, j \in \mathcal{I}$,*

$$f_i - f_j - \langle g_j, x_i - x_j\rangle - \frac{1}{2L(\delta)}\|g_i - g_j\|^2 + \delta \ge 0, \quad \forall\delta \ge 0. \tag{3.2}$$

Conversely, if the above condition holds then $\mathcal{H}$ is $c_L L(\cdot/c_L)$-interpolable.

Recall our motivating example of (β, p)-Hölder smooth functions always satisfy an inexactly smooth bound with $L(\delta) = \left(\frac{1-p}{1+p}\frac{1}{2\delta}\right)^{\frac{1-p}{1+p}}\beta^{\frac{2}{1+p}}$ by Proposition 2.1. Hence, from (3.1), c_L is exactly the maximum of $|1-s|^{(p+1)} - s^{(p+1)} + (p+1)\,s^p$ for $s \ge 0$, which is uniformly upper bounded by $c_L \le 2^{1-p} \le 2$ where the first inequality above is a direct application of the scalar ratio bound (3.4) derived in the proof of Theorem 3.2.[1]

[1] For (β, p)-Hölder smooth functions, $\theta_L(s) = \frac{\beta}{p+1}|s|^{p+1}$ and the supremum of 2^{1-p} in (3.4) is attained when $r = -s$.

This is tight when $p = 1$ but may be slack for $p < 1$. Note that the previous example (2.2) established that any interpolation theorem based on inexact quadratic bounds will be slack by at least a factor of $\left(\frac{p+1}{2p}\right)^p$. Combined, these give the following interpolation conditions for Hölder smooth functions, where we take the infimum over $\delta \geq 0$ in (3.2). Note that the sufficiency side can be tightened with the exact value of c_L instead of the bound 2^{1-p} given above.

Corollary 3.4. *If a set of observations $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{I}}$ is interpolable by some convex, (β, p)-Hölder smooth function for $p \in (0, 1]$, then for each $i, j \in \mathcal{I}$,*

$$f_i - f_j - \langle g_j, x_i - x_j\rangle - \beta^{-1/p}\frac{p}{p+1}\|g_i - g_j\|^{(p+1)/p} \geq 0. \tag{3.3}$$

Conversely, if the above condition holds then $\mathcal{H}$ is interpolable by a convex, $\left(2\left(\frac{p+1}{4p}\right)^p\beta, p\right)$-Hölder smooth function.

Considering the limiting case as $p \to 0$, the condition above becomes

$$f_i - f_j - \langle g_j, x_i - x_j\rangle - \iota_{[0,\beta]}(\|g_i - g_j\|) \geq 0$$

where $\iota_{[0,\beta]}$ is the indicator for the interval $[0, \beta]$. Hence the condition requires $f_i \geq f_j + \langle g_j, x_i - x_j\rangle$ and $\|g_i - g_j\| \leq \beta$. However, in this $p = 0$ special case, despite having $c_L = 2$, these are necessary and sufficient for $(\beta, 0)$-interpolation.

Theorem 3.5. *Consider any set of observations $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{I}}$. Then $\mathcal{H}$ is interpolable by a $(\beta, 0)$-Hölder smooth, convex function if and only if $\mathcal{H}$ is $L(\delta) = \frac{\beta^2}{2\delta}$-interpolable if and only if for all $i, j \in \mathcal{I}$,*

$$f_i \geq f_j + \langle g_j, x_i - x_j\rangle \qquad \text{and} \qquad \|g_i - g_j\| \leq \beta.$$

Since $p = 0$ admits a sharper exact interpolation theorem, further improvements for $p \in (0, 1)$ may be possible. However, the previous example in (2.2) establishes that a gap of at least $\left(\frac{p+1}{2p}\right)^p > 1$ is necessary for any approach for Hölder smoothness with $p \in (0, 1)$ based on inexact smoothness. As a result, Corollary 3.4's condition is only necessary and sufficient at $p = 1$ and $p = 0$.

3.1 Proof of Interpolation Results

In this section, we prove the above interpolation theory, utilizing the calculus results built in Section 2.2. The primary work in proving Theorem 3.3 and Theorem 3.5 is a construction of the necessary interpolating function given the claimed interpolation conditions.

Proof of Theorem 3.2. The lower bound $c_L \geq 1$ is immediate. Set $s = 0$, and recall that $\theta_L(0) = 0$ with $0 \in \partial\theta_L(0)$ by Lemma 2.9. Notice that $c \geq 1$ since the defining inequality in (3.1) gives

$$\theta_L(r) \leq c\theta_L(r), \quad \forall r.$$

The rest of the proof proceeds in three steps. We first prove a general implication by considering

1. $\|g_x - g_y\| \leq \|s_{xy}\|, \quad \forall x, y$ with $g_x \in \partial f(x), g_y \in \partial f(y),$ and $s_{xy} \in \partial(\theta_L \circ \|\cdot\|)(x - y)$,
2. $f(y) \leq f(x) + \langle g_x, y - x\rangle + \theta_L(\|y - x\|), \quad \forall x, y$ with $g_x \in \partial f(x)$,

and demonstrating that the first condition implies the second. Next, we specialize to $f = \theta_L$ and bound the self-smoothness constant in (3.1) with the scalar ratio below

$$c_L \leq \sup_{r \neq s} \left\{ \frac{|g_r - g_s|}{|g_{rs}|} : g_r \in \partial\theta_L(r), g_s \in \partial\theta_L(s), g_{rs} \in \partial\theta_L(r-s) \right\}. \tag{3.4}$$

Finally, we complete the proof by establishing a universal bound of 2 on this quantity.

We first show (1. $\implies$ 2.). Fix x, y and arbitrary $g_x \in \partial f(x)$. Letting $g_t \in \partial f(x + t(y-x))$ and $s_t \in \partial(\theta_L \circ \|\cdot\|)(t(y-x))$ be arbitrary for any $t \in [0,1]$, we note that

$$\begin{aligned} f(y) - f(x) - \langle g_x, y - x\rangle &= \int_0^1 \langle g_t - g_x, y - x\rangle \, dt \\ &\leq \int_0^1 \|g_t - g_x\| \|y - x\| dt \\ &\leq \int_0^1 \|s_t\| \|y - x\| dt \\ &= \theta_L(\|y - x\|) - \theta_L(0) = \theta_L(\|y - x\|), \end{aligned} \tag{3.5}$$

where the nonsmooth fundamental theorem of calculus [18, Theorem 2.3.4] implies the first equality, Cauchy-Schwarz implies the first inequality, and the hypothesis implies the second inequality. The last equality follows from two nonsmooth calculus rules. First, the nonsmooth chain rule [1, Corollary 16.72] applied to $(\theta_L \circ \|\cdot\|)$ gives $\|s_t\|\|y - x\| = \langle s_t, y - x\rangle$. Second, we consider the fundamental theorem of calculus again applied to $\theta_L \circ \|\cdot\|$ along the segment from 0 to $y - x$. Rearranging yields $f(y) \leq f(x) + \langle g_x, y - x\rangle + \theta_L(\|y - x\|)$.

We now construct the bound as in (3.4). By the definition of c_L in (3.1), for all $r, s \in \mathbb{R}$ with $g_s \in \partial\theta_L(s)$ the following bound holds

$$\theta_L(r) \leq \theta_L(s) + g_s \cdot (r - s) + c_L \theta_L(r - s).$$

Suppose the following bound holds for some fixed $c > 0$,

$$|g_r - g_s| \leq c|g_{rs}|, \quad \forall r, s \in \mathbb{R},\ g_r \in \partial\theta_L(r),\ g_s \in \partial\theta_L(s),\ g_{rs} \in \partial\theta_L(r - s).$$

By applying (1. $\implies$ 2.) as above, it holds that

$$\theta_L(r) \leq \theta_L(s) + g_s \cdot (r - s) + c\theta_L(r - s).$$

Since this bound holds trivially if $r = s$ (as $\theta_L(0) = 0$ by Lemma 2.9), we can bound c_L as in (3.4).

Finally, we bound this value by 2. Without loss of generality, suppose $r > s$. Recall that θ_L is convex, even, and minimized at 0. Respectively, the subdifferential $\partial\theta_L$ is monotone, $\partial\theta_L(s) = -\partial\theta_L(-s)$, and $0 \in \partial\theta_L(0)$.

We first consider the case where $r \geq 0 \geq s$. Here $|r - s| = r + |s|$, so $g_{rs} \in \partial\theta_L(r + |s|)$. Monotonicity of the subdifferential with evenness of θ_L gives $|g_r| \leq |g_{rs}|$ and $|g_s| \leq |g_{rs}|$. Since θ_L minimizes and vanishes at 0 (see Lemma 2.9), $g_{rs} \geq 0$ as well. Hence

$$|g_r - g_s| \leq |g_r| + |g_s| \leq 2g_{rs}.$$

To conclude, we consider the case where $r > s > 0$. Recall $\phi_L(s, \delta) = \frac{L(\delta)}{2}s^2 + \delta$ and denote $\delta_\star(s) = \operatorname{argmin}_{\delta \geq 0} \phi_L(s, \delta)$. Note that by Lemma 2.8, both $\partial\theta_L(s)$ and $\partial\theta_L(r)$ are singletons:

$$\partial\theta_L(s) = \{sL(\delta_\star(s))\}, \quad \partial\theta_L(r) = \{rL(\delta_\star(r))\}.$$

Adding the two inequalities $\phi_L(s, \delta_\star(s)) \leq \phi_L(s, \delta_\star(r))$ and $\phi_L(r, \delta_\star(r)) \leq \phi_L(r, \delta_\star(s))$ and simplifying terms that cancel,

$$(r^2 - s^2)(L(\delta_\star(r)) - L(\delta_\star(s))) \leq 0 \Longrightarrow L(\delta_\star(r)) \leq L(\delta_\star(s)).$$

Therefore

$$\begin{aligned} 0 \leq g_r - g_s &= rL(\delta_\star(r)) - sL(\delta_\star(s)) \\ &= (r-s)L(\delta_\star(r)) + s(L(\delta_\star(r)) - L(\delta_\star(s))) \\ &\leq (r-s)L(\delta_\star(r-s)) = g_{rs}, \end{aligned}$$

where we utilize monotonicity of the subgradient in the first inequality and monotonicity of $L(\delta_\star(\cdot))$ in the second. The remaining case $0 > r > s$ follows by applying the above to $(-s, -r)$ and invoking evenness. With all cases exhausted, we may bound (3.4) by 2. □

Proof of Theorem 3.3. Suppose $f(x)$ is an $L(\cdot)$-inexactly smooth function which interpolates the observations $\mathcal{H}$. By the cocoercive-type condition 1. from Lemma 2.4 with $x = x_j$ and $y = x_i$, the exact claimed interpolation conditions hold for each $i, j \in \mathcal{I}$.

To show the other direction, we define our interpolation as the function $r = (\max_{i \in \mathcal{I}} r_i^*)^*$ where

$$r_i(x) = f_i + \langle g_i, x - x_i \rangle + \theta_L(\|x - x_i\|).$$

Calculating the conjugates of each r_i, one has $r_i^*(g) = -f_i + \langle x_i, g \rangle + \theta_L^*(\|g - g_i\|)$.

This claimed interpolation r is convex as all conjugates are convex. Our calculus rules allow us to verify its inexact smoothness. Each individual r_i is $c_L L(\cdot/c_L)$-inexactly smooth by the sum rule in Lemma 2.6 since the linear term $f_i + \langle g_i, x - x_i \rangle$ does not affect the smoothness and $\theta_L(\|x - x_i\|)$ is $c_L L(\cdot/c_L)$-inexactly smooth by Lemma 2.7. Then the $c_L L(\cdot/c_L)$-inexact smoothness of r follows from the conjugate maximum formula in Lemma 2.6.

All that remains is to verify that r interpolates the given first-order information. Note that $r(x_j) = f_j$ and $g_j \in \partial r(x_j)$ if and only if $\max_{i \in \mathcal{I}} r_i^*(g_j) = -f_j + \langle x_j, g_j \rangle$ and $x_j \in \partial(\max_{i \in \mathcal{I}} r_i^*)(g_j)$. The key step in establishing this is showing that the interpolation conditions guarantee that r_j^* attains this maximum at g_j. To see this, observe that for any $i \in \mathcal{I}$,

$$\begin{aligned} r_j^*(g_j) &= -f_j + \langle x_j, g_j \rangle \\ &\geq -f_i + \langle x_i, g_j \rangle + \theta_L^*(\|g_i - g_j\|) \\ &= r_i^*(g_j), \end{aligned}$$

where the inequality is our interpolation condition between i and j (taking the supremum over $\delta \geq 0$). From this, the fact that r interpolates follows since $r^*(g_j) = -f_j + \langle x_j, g_j \rangle$ and $x_j \in \partial r_j^*(g_j) \subseteq \partial(\max_{i \in \mathcal{I}} r_i^*)(g_j)$. □

Proof of Theorem 3.5. Recall by Proposition 2.1 that a function f is $(\beta, 0)$-Hölder smooth if and only if it is $L(\delta) = \frac{\beta^2}{2\delta}$-inexactly smooth. Therefore, the existence of a $(\beta, 0)$-Hölder smooth function interpolating the observations $\mathcal{H}$ is equivalent to $L(\delta) = \frac{\beta^2}{2\delta}$ interpolation.

Supposing there exists an $L(\delta) = \frac{\beta^2}{2\delta}$-inexactly smooth function interpolating these observations, by Theorem 3.3, it must hold for all $i, j \in \mathcal{I}$ and $\delta \geq 0$ that $f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{\delta}{\beta^2}\|g_i - g_j\|^2 + \delta \geq 0$. Taking the infimum over $\delta \geq 0$, observe that the interpolation condition above is then equivalent to $f_i \geq f_j + \langle g_j, x_i - x_j \rangle$ and $\|g_i - g_j\| \leq \beta$.

Conversely, consider the convex function $h(x) = \max\{f_i + \langle g_i, x - x_i \rangle\}$. Since $\partial h \subseteq \operatorname{conv}\{g_i\}$, for any $g, g' \in \partial h$ with $g = \sum \alpha_i g_i$ and $g' = \sum \gamma_j g_j$, $(\beta, 0)$-Hölder smooth follows as $\|g - g'\| \leq \beta$ by convexity of the norm. Finally, since $f_i \geq f_j + \langle g_j, x_i - x_j \rangle$, the ith component of h is active at x_i, and therefore $h(x_i) = f_i$ and $g_i \in \partial h(x_i)$. Hence, h interpolates the observations. □

3.2 Inexactly Smooth Convex PEPs

Our interpolation theorems enable the design of PEPs [10, 29, 30] for inexactly smooth problems. PEPs provide a structured way to analyze N-step fixed-step first-order methods (FSFOM). These methods, parameterized by a lower triangular matrix $W \in \mathbb{R}^{N\times N}$, are defined by the iteration

$$x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$$

for $n = 1, \dots, N$, where $g_i \in \partial f(x_i)$ are subgradients computed at each iteration. For ease of presentation, the rows of W are indexed from 1 to N while the columns are indexed from 0 to $N-1$.

Consider minimizing an $L(\cdot)$-inexactly smooth, convex function f with a minimizer $x_\star$ satisfying $\|x_0 - x_\star\| \le D$. For an algorithm defined by some W, the Performance Estimation Problem finds the worst-case final objective $f(x_N) - f(x_\star)$ over all possible problem instances. Formally,

$$p_{true} = \begin{cases} \max\limits_{f,\, x_0 \in \mathbb{R}^d} & f(x_N) - f(x_\star) \\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i \quad \forall n = 1, 2, \dots, N \\ & \|x_0 - x_\star\|^2 \le D^2 \\ & f \text{ is } L(\cdot)\text{-inexactly smooth, convex, and minimized at } x_\star. \end{cases} \tag{3.6}$$

The key observation to tractably solve PEPs is that the algorithm's trajectory and performance depend only on the first-order information at the iterates $x_0, \dots, x_N$ and at the minimizer $x_\star$. For ease in referring to such points, we consider the index set $\mathcal{I}_N^\star = \{0, 1, \dots, N, \star\}$. Then the relevant quantities to the performance of the algorithm parameterized by W are the values

$$x_i, \qquad f_i = f(x_i), \quad g_i \in \partial f(x_i) \qquad \forall i \in \mathcal{I}_N^\star$$

where we require $g_\star = 0$ at the minimizer $x_\star$. From our interpolation theorem (Theorem 3.3), we know that the following must be nonnegative for all $i, j \in \mathcal{I}_N^\star$ and all $\delta_{i,j} \ge 0$

$$\begin{aligned} \mathcal{D} &= D^2 - \|x_0 - x_\star\|^2 \ge 0 \\ \mathcal{Q}_{i,j,\delta_{i,j}} &= f_i - f_j - \langle g_j, x_i - x_j\rangle - \frac{1}{2L(\delta_{i,j})}\|g_i - g_j\|^2 + \delta_{i,j} \ge 0. \end{aligned}$$

Note this can be reduced to a finite set of inequalities by taking a supremum over $\delta_{i,j}$, resulting in a compressed nonnegative condition $\mathcal{Q}_{i,j} = f_i - f_j - \langle g_j, x_i - x_j\rangle - \theta_L^*(\|g_i - g_j\|) \ge 0$. Using the sufficiency side of Theorem 3.3, we know an $L(\delta)$ inexactly smooth instance exists agreeing with the observed first-order information whenever one has $\mathcal{D} \ge 0$ and for all $i, j \in \mathcal{I}_N^\star$ and all $\delta_{i,j} \ge 0$

$$\mathcal{Q}_{i,j,\delta_{i,j}}^{c_L} = f_i - f_j - \langle g_j, x_i - x_j\rangle - \frac{c_L}{2L(c_L\delta_{i,j})}\|g_i - g_j\|^2 + \delta_{i,j} \ge 0.$$

Again, we define a compressed notation as $\mathcal{Q}_{i,j}^{c_L} = f_i - f_j - \langle g_j, x_i - x_j\rangle - \theta^*_{L(c_L\cdot)/c_L}(\|g_i - g_j\|) \ge 0$.

From these necessary and sufficient conditions, we can define closely related optimization problems over finitely many variables. An upper bound on p_{true} is provided by

$$p_{true} \le p_{interp} = \begin{cases} \max\limits_{x,f,g \in (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d)^{N+2}} & f_N - f_\star \\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i \quad \forall n = 1, 2, \dots, N \\ & g_\star = 0 \\ & \mathcal{D} \ge 0 \\ & \mathcal{Q}_{i,j} \ge 0 \qquad \forall i, j \in \mathcal{I}_N^\star. \end{cases} \tag{3.7}$$

We denote the same problem with stricter $\mathcal{Q}_{i,j}^{c_L} \geq 0$ constraints by $p_{interp}^{c_L}$. This also provides a lower bound for p_{true}.

The standard approach to further simplifying such PEP reformulations is to consider a Gram change of variables. Define $P = [x_0 - x_\star \mid g_0 \mid g_1 \mid \ldots \mid g_N]$ and the Gram matrix of all inner products between these vectors as $G = P^T P$. After substituting the equality definitions of $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$ and $g_\star = 0$ into the remaining inequality constraints, observe that $\mathcal{D}$ and $\mathcal{Q}_{i,j}$ are linear in $F = [f_\star, f_0, \ldots, f_N]^T$ and concave in G (to see this, note $\theta_L^* \circ \sqrt{\cdot}$ is convex since it takes the form of a supremum of affine functions). We denote these concave functions by $\mathcal{D}(F,G)$ and $\mathcal{Q}_{i,j}(F,G)$. Changing over to these as the variables makes all of the above constraints convex. Additionally, as a Gram matrix, we require G to be positive semidefinite. We denote this by

$$p_{interp} \leq p_{gram} = \begin{cases} \max\limits_{F,G} & f_N - f_\star \\ \text{s.t.} & \mathcal{D}(F,G) \geq 0 \\ & \mathcal{Q}_{i,j}(F,G) \geq 0 \qquad \forall i,j \in \mathcal{I}_N^\star \\ & G \succeq 0. \end{cases} \tag{3.8}$$

Provided $d \geq N+2$, one can factor G to recover P, making this Gram reformulation exact.

The above PEP formulation (3.8) is a convex problem in F, G. The Lagrange dual problem certifies upper bounds on $f_N - f_\star$ via nonnegative weights $\nu, \lambda_{i,j}$ for the inequality constraints and a positive semidefinite Z. We denote the dual by d_{gram}. The following theorem relates all of the defined PEP problems, establishing, in particular, that strong duality holds above under a mild regularity condition on W.

Theorem 3.6. *Let $L(\cdot)$ satisfy our assumptions and suppose $d \geq N+2$. Given $W_{i,i-1} \neq 0$ for all $i = 1, \ldots, N$, strong duality holds between p_{gram} and d_{gram}. Therefore,*

$$p_{gram}^{c_L} = d_{gram}^{c_L} = p_{interp}^{c_L} \leq p_{true} \leq p_{interp} = p_{gram} = d_{gram}.$$

3.3 Proof of Theorem 3.6

First, we show that for any $L(\cdot)$ satisfying our assumptions, we can construct a Huber-type function that is convex and $L(\cdot)$-inexactly smooth in an arbitrary dimension d. From this, it follows that p_{gram} (3.8) has a Slater point, given W has nonzero diagonal. We defer the proofs of these two results, Lemma 3.7 and Lemma 3.8, to Appendix B.

Lemma 3.7. *Let $L(\cdot)$ satisfy our assumptions, and choose any $R > 0$. Then the function*

$$h(x) = \begin{cases} \frac{k}{2}\|x\|^2 & \|x\| \leq R \\ kR\|x\| - \frac{kR^2}{2} & \|x\| > R \end{cases}, \quad k := \inf_{\delta \geq 0} \left\{ \frac{L(\delta) + \sqrt{L(\delta)^2 + 2\delta L(\delta)/R^2}}{2} \right\}$$

is $L(\cdot)$-inexactly smooth and $k > 0$. Furthermore, for any positive definite Q, with $\|Q\|_{\text{op}} \leq 1$,

$$f(x) := h(Q^{1/2}x) = \begin{cases} \frac{k}{2}\|x\|_Q^2 & \|x\|_Q \leq R \\ kR\|x\|_Q - \frac{kR^2}{2} & \|x\|_Q > R \end{cases}$$

is $L(\cdot)$-inexactly smooth as well, where $\|x\|_Q := \sqrt{x^T Q x}$.

Lemma 3.8. *Let $L(\cdot)$ satisfy our assumptions. Under the assumption that $W_{i,i-1} \neq 0$ for each $i = 1, \ldots, N$, there exists feasible F, G for the convex problem p_{gram} (3.8) such that $G \succ 0$ and $\mathcal{Q}_{i,j}(F,G) > 0$ for all $i \neq j \in \mathcal{I}_N^\star$.*

Existence of a Slater point from the above Lemma 3.8 ensures strong duality between the convex primal and dual formulations p_{gram} and d_{gram} (see [28, Theorem 28.2]). Similarly, the lower bounding formulations of $p_{gram}^{c_L}$ and $d_{gram}^{c_L}$ are equal. Then, recalling that if $d \geq N+2$, one may always factor positive semidefinite G into its Cholesky form, the Gram reformulations are exact, $p_{interp} = p_{gram}$ and $d_{gram}^{c_L} = p_{interp}^{c_L}$. Finally, the inequalities $p_{interp}^{c_L} \leq p_{true} \leq p_{interp}$ follow from our interpolation theorem (Theorem 3.3).

4 A Constructive Approach to Optimized Algorithm Design

The above interpolation and performance estimation theory directly enables optimized algorithm designs. In particular, the constructive approach of Drori and Taylor [9] can be readily generalized. Below, we begin by discussing minimax optimality among algorithms and formalizing their constructive PEP approach to algorithm design in the context of inexactly smooth functions. Then we describe three resulting designs targeting different levels of generality in the inexact smoothness function $L(\cdot)$. These results are briefly summarized as:

Section 4.2 presents an exactly minimax optimal method for $(\beta, 0)$-Hölder smooth functions (i.e., $L(\delta) = \frac{\beta^2}{2\delta}$-inexactly smooth). Section 4.3 presents an optimized method for any (β, p)-Hölder smooth functions by optimizing performance over $L(\delta) = \kappa/\delta^q$-inexactly smooth functions. This method is big-O optimal and we conjecture that it possesses the optimal leading coefficient. Finally, Section 4.4 presents a universal, parameter-free method for any $L(\cdot)$, generalizing Nesterov's Universal Fast Gradient Method (UFGM) [26] and the Optimized Backtrackable Linesearch method (OBL) [27]. The resulting Universal Optimized Backtrackable Linesearch method (UOBL) offers optimized big-O optimal guarantees for any Hölder smooth setting or sums thereof.

Next, we formally define minimax optimality and big-O optimality. Optimality in algorithm design is defined in terms of a considered family of problem instances and a family of algorithms. As problems, given a function $L(\cdot)$ and some $D > 0$, we consider minimizing $L(\cdot)$-inexactly smooth convex functions from an initialization x_0 with $\|x_0 - x_\star\| \leq D$. Denote this class

$$\begin{aligned} \mathtt{P}_{L(\cdot),D} := \{(f, x_0) \colon & f \text{ is convex, } L(\cdot)\text{-inexactly smooth,} \\ & \text{attains a minimizer at } x_\star \text{ with } \|x_0 - x_\star\| \leq D, \\ & \text{and a subgradient oracle } g \text{ such that } g(x) \in \partial f(x)\}. \end{aligned} \tag{4.1}$$

Note that $\mathtt{P}_{L(\cdot),D}$ contains problems over every dimension d. As algorithms, we consider N-step methods constructing points $x_1, \ldots, x_N$ satisfying the subgradient span condition $x_n \in x_0 + \operatorname{span}\{g_0, \ldots, g_{n-1}\}$ with $g_i = g(x_i)$. Denote the set of such methods by $\mathtt{A}_{\text{span}}$. Note that this contains, for example, all FSFOM studied in Section 3.2.

The task of finding the algorithm with the best worst-case performance against a family of problem instances, measured by final objective gap, is then

$$\min_{\mathtt{a} \in \mathtt{A}_{\text{span}}} \max_{(f,x_0) \in \mathtt{P}_{L(\cdot),D}} f(x_N) - f(x_\star). \tag{4.2}$$

We say an algorithm $\mathtt{a}$ is minimax optimal if it attains the above min. It is big-O optimal if for all N, it remains within a constant factor of attaining this rate. Proving minimax optimality of methods historically has been done by considering the dual maximin problem, seeking a hard problem instance for all algorithms [24]. By weak duality, (4.2) is lower bounded by

$$\max_{(f,x_0) \in \mathtt{P}_{L(\cdot),D}} \min_{\mathtt{a} \in \mathtt{A}_{\text{span}}} f(x_N) - f(x_\star). \tag{4.3}$$

Strong duality often holds. For constant $L(\cdot) = L$, this is established by the OGM method [20] and matching lower bound [8]. For $L(\delta) = \kappa/\delta$, our Theorem 4.2 below establishes strong duality.

4.1 Extension of the Constructive Approach of [9]

Here we consider the constructive approach to algorithm design problems with semidefinite programming PEPs of Drori and Taylor [9]. These techniques generalize directly to the inexactly smooth setting with its convex PEP formulation. To this end, consider the following hypothetical first-order method, iterating

$$\begin{aligned} &x_n \in \operatorname{argmin}\{f(x) \colon x \in x_0 + \operatorname{span}\{g_0, g_1, \ldots, g_{n-1}\}\} \\ &g_n \in \partial f(x_n) \text{ such that } \langle g_n, g_i\rangle = 0,\ \forall 0 \le i < n. \end{aligned}$$

For the sake of this motivation, assume the above argmin is nonempty and consider any selection of (possibly adversarial) x_n and g_n. Following the nomenclature of [9], we refer to such a hypothetical algorithm as a Greedy First-Order Method (GFOM).

The constructive approach to algorithm design then proceeds by (i) solving the PEP problem associated with GFOM, (ii) computing dual multipliers proving its convergence rate, and (iii) identifying a fixed-step first-order method with the same performance (and, in fact, the same PEP proof) as GFOM. Formally, denote the PEP for GFOM as

$$p_{true}^{\texttt{alg}} = \begin{cases} \max_{(f,x_0)} & f(x_N) - f_\star \\ \text{s.t.} & x_n \text{ is constructed by some GFOM} \\ & (f, x_0) \in \mathtt{P}_{L(\cdot),D.} \end{cases} \tag{4.4}$$

Then, utilizing our interpolation theory, we derive the following upper bounding problem

$$p_{true}^{\texttt{alg}} \le p_{interp}^{\texttt{alg}} = \begin{cases} \max_{x_i, f_i, g_i} & f_N - f_\star \\ \text{s.t.} & \langle g_i, g_j\rangle = 0, \quad \forall 0 \le j < i = 1, \ldots, N \\ & \langle g_i, x_j - x_0\rangle = 0, \quad \forall 1 \le j \le i = 1, \ldots, N \\ & g_\star = 0 \\ & \mathcal{D} \ge 0 \\ & \mathcal{Q}_{i,j} \ge 0, \quad i,j \in \mathcal{I}_N^\star. \end{cases} \tag{4.5}$$

Observe that after substituting the orthogonality constraints, this is a convex problem in the variables f_i, $\langle g_i, x_j\rangle$, and $\|g_i\|^2$. The corresponding dual problem can be formulated as follows (the derivation of this program is deferred to Appendix C)

$$d_{interp}^{\texttt{alg}} = \begin{cases} \min_{\lambda,t,s} & \frac{1}{2}D^2 s + \sum_{i,j\in\mathcal{I}_N^\star} \lambda_{i,j} L^{\leftarrow}(\lambda_{i,j}/t_{i,j}) \\ \text{s.t.} & \lambda_{\star,j}^2/s \le \sum_{i=0}^{j-1} t_{i,j} + \sum_{i=j+1}^{N} t_{j,i} + t_{\star,j} - t_{j,\star}, \quad \forall j = 0, \ldots, N \\ & \sum_{i=j+1}^{N} \lambda_{j,i} - \sum_{i=0}^{j-1} \lambda_{i,j} = \lambda_{\star,j} - \lambda_{j,\star}, \quad \forall j \ne N \\ & \sum_{i=0}^{N-1} \lambda_{i,N} = \sum_{i=0}^{N-1} \lambda_{\star,i} - \lambda_{i,\star} \\ & \sum_{i=0}^{N} \lambda_{\star,i} - \lambda_{i,\star} = 1 \\ & \lambda \ge 0,\ t \ge 0,\ s \ge 0. \end{cases} \tag{4.6}$$

Recall $sL^{\leftarrow}(s)$ is convex by our assumption on $L(\cdot)$ and Lemma 2.2. From this, the convexity of the dual objective is clear as it uses the perspective function of this [2, Section 3.2.6].

In particular, solving this dual problem gives certificates λ, t, s that prove a convergence rate for GFOM. The "Subspace Search Elimination Procedure" (SSEP) of [9] shows how to construct an FSFOM for which this same certificate also proves a convergence rate. Namely, after a notational rearrangement of [9, Corollary 1], set $z_0 = x_0$, $z_1 = x_0 - \lambda_{\star,0} g_0$, and iterate for $n = 1, \dots, N$

$$\begin{aligned} x_n &= \frac{\sum_{i=0}^{n-1} (\lambda_{i,n} x_i - t_{i,n} g_i) + \lambda_{\star,n} z_n}{\sum_{i=0}^{n-1} \lambda_{i,n} + \lambda_{\star,n}} \\ z_{n+1} &= z_n - \lambda_{\star,n} g_n. \end{aligned} \tag{4.7}$$

The following lemma explicitly connects this to the quantities of Drori and Taylor and extracts the resulting convergence guarantee.

Lemma 4.1. *Fix $N \geq 1$, $D > 0$ and let (λ, t, s) be feasible for (4.6). Suppose $\Lambda_n := \sum_{i=0}^{n-1} \lambda_{i,n} + \lambda_{\star,n} \neq 0$ for all $n = 1, \dots, N$. Then the method (4.7) is exactly the fixed-step method of [9, Corollary 1] under the identifications*

$$\begin{aligned} \tilde{\gamma}_{n,n} = \Lambda_n, \quad \tilde{\gamma}_{n,j} = -\lambda_{j,n} \qquad & \forall j = 1, \dots, n-1, \\ \tilde{\beta}_{n,j} = t_{j,n} + \lambda_{\star,n} \lambda_{\star,j} \qquad & \forall j = 0, \dots, n-1. \end{aligned} \tag{4.8}$$

Further, for any problem instance $(f, x_0) \in \mathtt{P}_{L(\cdot),D}$, this method inherits the dual objective bound

$$f(x_N) - f_\star \leq \frac{1}{2} D^2 s + \sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j} L^{\leftarrow}(\lambda_{i,j}/t_{i,j}).$$

Proof. Recall that we set $z_0 = x_0$, $z_1 = x_0 - \lambda_{\star,0} g_0$ and, for $n = 1, \dots, N$,

$$x_n = \frac{\sum_{i=0}^{n-1} \lambda_{i,n} x_i + \lambda_{\star,n} z_n - \sum_{i=0}^{n-1} t_{i,n} g_i}{\Lambda_n}, \qquad z_{n+1} = z_n - \lambda_{\star,n} g_n. \tag{4.9}$$

Unrolling the z-recurrence gives $z_n = x_0 - \sum_{k=0}^{n-1} \lambda_{\star,k} g_k$. Substituting this into (4.9) yields

$$\Lambda_n x_n = \sum_{i=0}^{n-1} \lambda_{i,n} x_i + \lambda_{\star,n} x_0 - \sum_{i=0}^{n-1} (t_{i,n} + \lambda_{\star,n} \lambda_{\star,i})\, g_i.$$

Subtracting $\Lambda_n x_0 = \sum_{i=0}^{n-1} \lambda_{i,n} x_0 + \lambda_{\star,n} x_0$ from both sides, canceling terms, and dividing by $\Lambda_n \neq 0$ gives

$$x_n = x_0 + \sum_{i=1}^{n-1} \frac{\lambda_{i,n}}{\Lambda_n} (x_i - x_0) - \sum_{i=0}^{n-1} \frac{t_{i,n} + \lambda_{\star,n} \lambda_{\star,i}}{\Lambda_n} g_i. \tag{4.10}$$

The correspondence given in (4.8) applied to (4.10) is exactly the fixed-step formula given in [9, Corollary 1]. Since $\tilde{\gamma}_{n,n} = \Lambda_n \neq 0$ by assumption, we may apply their results to derive the worst-case bound $f(x_N) - f_\star \leq \frac{1}{2} D^2 s + \sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j} L^{\leftarrow}(\lambda_{i,j}/t_{i,j})$, inherited from GFOM. □

4.2 An Exactly Optimal Method for $(\beta, 0)$-Hölder Smooth Convex Minimization

As a first application, we apply the constructive approach to $(\beta, 0)$-Hölder smooth convex minimization. Such problems have bounded differences between subgradients, i.e., for all x, y with $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$, one has $\|g_x - g_y\| \leq \beta$. By Proposition 2.1, this is exactly the class of $L(\delta) = \frac{\beta^2}{2\delta}$-inexactly smooth convex functions. Our interpolation theorem is tight for this setting, by

Theorem 3.5. Hence from Theorem 3.6, we have $p_{true} = p_{interp}$. Similarly, the GFOM PEP here satisfies $p_{true}^{\texttt{alg}} = p_{interp}^{\texttt{alg}}$.

This class is closely related to the well-studied model of M-Lipschitz convex minimization. While any M-Lipschitz convex function is necessarily $(2M, 0)$-Hölder smooth, $(\beta, 0)$-Hölder smoothness combined with existence of a minimizer having a zero subgradient implies β-Lipschitzness. As a result, these distinct models are equivalent only up to differences in universal constants. Which of these two models is more relevant is a modeling choice; as one nice property, $(\beta, 0)$-Hölder smoothness is tilt invariant (i.e., preserved under addition with linear functions), making its corresponding dual property translation invariant.

A minimax optimal method for Lipschitz convex problems was derived using the above constructive approach in its original development [9]. Applying this framework to $(\beta, 0)$-Hölder smooth convex problems below yields a similar (but distinct) minimax optimal method here.

4.2.1 A Minimax Optimal Algorithm With $L(\delta) = \frac{\beta^2}{2\delta}$, the dual program (4.6) simplifies to

$$d_{interp}^{\texttt{alg}} = \begin{cases} \min_{\lambda, t \geq 0} & \frac{D^2 + \sum_{0 \leq i < j \leq N} \beta^2 t_{i,j}}{2\left(\sum_{i=0}^{N} \lambda_{\star,i} - \lambda_{i,\star}\right)} \\ \text{s.t.} & \sum_{i=j+1}^{N} \lambda_{j,i} - \sum_{i=0}^{j-1} \lambda_{i,j} = \lambda_{\star,j} - \lambda_{j,\star}, \quad \forall j \neq N \\ & \sum_{i=0}^{N-1} \lambda_{i,N} = \sum_{i=0}^{N-1} \lambda_{\star,i} - \lambda_{i,\star} \\ & \sum_{i=0}^{j-1} -t_{i,j} + \sum_{i=j+1}^{N} -t_{j,i} - t_{\star,j} + t_{j,\star} + \lambda_{\star,j}^2 \leq 0, \quad \forall j = 0, \ldots, N. \end{cases}$$

A simple calculation verifies that the following is a feasible solution, setting

$$\begin{gathered} \lambda_{\star,i} = \frac{D\sqrt{2}}{\beta\sqrt{N+1}}, \quad i = 0, \ldots, N, \quad \lambda_{i,i+1} = \frac{\sqrt{2}D(i+1)}{\beta\sqrt{N+1}}, \quad i = 0, \ldots, N-1, \\ t_{i,j} = \frac{2D^2}{\beta^2 N(N+1)}, \quad i < j = 1, \ldots, N \end{gathered} \tag{4.11}$$

and all other variables as zero. The dual objective value of this candidate solution is

$$\frac{D^2 + \sum_{i,j} \beta^2 t_{i,j}}{2\sum_{i=0}^{N} \lambda_{\star,i}} = \frac{\beta D}{\sqrt{2(N+1)}}.$$

Algorithm 1 presents the algorithm induced by this certificate, reformulating the recurrence (4.7).

Algorithm 1 SSEP Method for $(\beta, 0)$-Hölder Smooth Convex Minimization

Input: $x_0 \in \mathbb{R}^d$, iteration budget N, parameters β, D

for $n = 1, \ldots, N$

$y_n = \frac{n}{n+1} x_{n-1} + \frac{1}{n+1} x_0$

$d_n = \frac{1}{n+1} \sum_{j=0}^{n-1} g_j$

$x_n = y_n - \frac{\sqrt{2}D\sqrt{N+1}}{\beta N} d_n$

Trivial modifications of the known hard instance for Lipschitz convex problems establish a matching lower bound via (4.3). The following theorem summarizes this minimax optimality.

Theorem 4.2. *For any $N \geq 1, D > 0$ and any convex, $(\beta, 0)$-Hölder smooth f with minimizer $x_\star$ satisfying $\|x_0 - x_\star\| \leq D$, Algorithm 1's terminal iterate is guaranteed to have*

$$f(x_N) - f(x_\star) \leq \frac{\beta D}{\sqrt{2(N+1)}}.$$

Moreover, this method is exactly minimax optimal for such minimization, i.e., solves (4.2).

Proof. The claimed convergence rate is immediate from Lemma 4.1. The lower bound follows from the following standard maximin hard instance design for nonsmooth optimization: Let $d = N + 1$ and e_i denote the ith standard basis vector. Consider $x_0 = 0$,

$$f(x) = \frac{\beta}{\sqrt{2}} \max\left\{ \max_{i=1,\dots,N+1} \langle x, e_i \rangle, -\frac{D}{\sqrt{N+1}} \right\}$$

and the subgradient oracle choosing

$$g(x) = \begin{cases} \frac{\beta}{\sqrt{2}} e_{i_\star} & f(x) > -\frac{\beta D}{\sqrt{2(N+1)}} \\ 0 & f(x) = -\frac{\beta D}{\sqrt{2(N+1)}} \end{cases}, \quad i_\star = \min\{i : f(x) = \beta \langle x, e_i \rangle / \sqrt{2}\}.$$

It is easy to verify that f is $(\beta, 0)$-Hölder smooth, convex, and has $\|x_0 - x_\star\| = D$ with minimizer at $x_\star = -\frac{D}{\sqrt{N+1}} \sum_{i=1}^{N+1} e_i$. Then the minimal objective value is $f(x_\star) = -\frac{\beta D}{\sqrt{2(N+1)}}$. A standard zero-chain argument (see [11, Theorem A.1]) establishes that any algorithm in $\mathtt{A}_{\text{span}}$ has $f(x_N) \geq 0$. Hence, every subgradient span method must have $f(x_N) - f(x_\star) \geq \frac{\beta D}{\sqrt{2(N+1)}}$ on this hard problem instance. Combined with our equal upper bound, this proves exact optimality. □

4.2.2 Algorithmic Performance Comparison with Lipschitz Problem Class Given $(\beta, 0)$-Hölder smoothness differs from Lipschitz continuity only up to small constants, here we briefly investigate the performance of algorithms across these classes. Namely, we consider three known optimal methods for M-Lipschitz convex minimization: (i) The subgradient method with constant stepsize $h = D/(M\sqrt{N+1})$ and final iterate averaging [3, Section 3.1], (ii) The subgradient method with nonconstant stepsizes but optimal final iterate [32], (iii) The SSEP induced method of [9]. Each of these methods has an equal worst-case performance of $MD/\sqrt{N+1}$.

Surprisingly, when numerically solving our PEP (3.8) for $(\beta, 0)$-Hölder smooth convex problems, these three methods appear to possess identical convergence rates for our setting of interest as well. See Fig. 1. This holds whether we set $M = \beta$ or heuristically set $M = \beta/\sqrt{2}$. In either case, these methods are strictly suboptimal, performing worse than Algorithm 1. With $M = \beta/\sqrt{2}$, the optimal methods for the Lipschitz setting appear to have the optimal asymptotic coefficient for the $(\beta, 0)$-Hölder smooth setting, making them only suboptimal in little-o terms.

4.3 Asymptotically Optimized Methods for Hölder Smooth Minimization

Next, we consider (β, p)-Hölder smooth functions with $p \in [0, 1]$. The limiting cases now have known optimal methods (the above method when $p = 0$ and OGM [20] when $p = 1$). We recall by Proposition 2.1 that if a function is (β, p)-Hölder smooth then it is $L(\delta) = \kappa/\delta^q$-inexactly smooth where $q = \frac{1-p}{1+p}$ and $\kappa = (q/2)^q \beta^{\frac{2}{1+p}}$. Recall that this is tight up to an absolute constant at most 1.263. Consequently, we study the class of $L(\delta) = \kappa/\delta^q$-inexactly smooth functions.

Numerically applying the constructive approach (4.6), we observed that regardless of the parameters N, κ, q, D, solutions exist with a structured sparsity pattern in λ. Specifically, only $\lambda_{n-1,n}$ and $\lambda_{\star,n}$ are nonzero. As a result, only the inequalities $\mathcal{Q}_{n-1,n}$ and $\mathcal{Q}_{\star,n}$ and respectively tolerances $\delta_{n-1,n}$ and $\delta_{\star,n}$ are needed to analyze the resulting methods. Once one fixes $\delta_{n-1,n}$ and $\delta_{\star,n}$ and the above sparsity pattern, the resulting λ values and induced algorithm are uniquely specified. One arrives at the following method, structurally identical to the Optimized Gradient Method of [20], stated in Algorithm 2, noting our convention $1/\infty = 0$ when applicable.

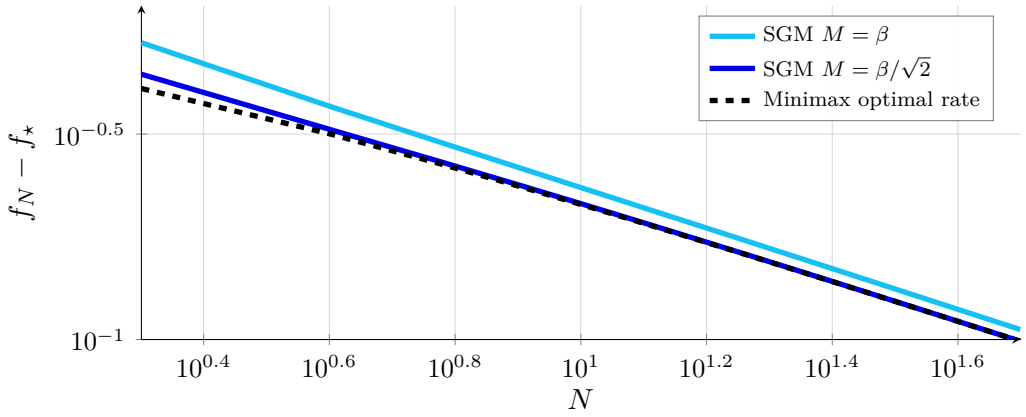

Figure 1: Suboptimality of the subgradient method [3, Section 3.1] on $(\beta, 0)$-Hölder smooth functions. Other optimal subgradient methods are omitted as they had numerically identical PEP values.

Algorithm 2 Inexactly Smooth Optimized Gradient Method

Input: x_0, iteration budget N, tolerance sequence $\{\delta_{n-1,n}\}, \{\delta_{\star,n}\}$

Initialize: $\tau_0 = \frac{1}{L(\delta_{0,1})} + \frac{1}{L(\delta_{\star,0})}$, $z_1 = x_0 - \tau_0 g_0$

for $n = 1, \dots, N$

$$\tau_n = \begin{cases} \tau_{n-1} + \dfrac{\frac{1}{L(\delta_{\star,n})} + \sqrt{\frac{1}{L(\delta_{\star,n})^2} + \frac{4\tau_{n-1}}{L(\delta_{n-1,n})}}}{2} & \text{if } n = N \\ \tau_{n-1} + \dfrac{\frac{1}{L(\delta_{n,n+1})} + \frac{1}{L(\delta_{\star,n})} + \sqrt{\left(\frac{1}{L(\delta_{n,n+1})} + \frac{1}{L(\delta_{\star,n})}\right)^2 + \frac{4\tau_{n-1}}{L(\delta_{n-1,n})} + \frac{4\tau_{n-1}}{L(\delta_{n,n+1})}}}{2} & \text{else} \end{cases}$$

$x_n = \frac{\tau_{n-1}}{\tau_n}\left(x_{n-1} - \frac{1}{L(\delta_{n-1,n})} g_{n-1}\right) + \frac{\tau_n - \tau_{n-1}}{\tau_n} z_n$

$z_{n+1} = z_n - (\tau_n - \tau_{n-1}) g_n$

All that remains is to pick a set of tolerances $\delta_{n-1,n}$ and $\delta_{\star,n}$, given $L(\delta) = \kappa/\delta^q$-inexact smoothness and $N, D > 0$. Alas, from numerical solutions to (4.6), an analytic formula remained elusive. However, as N grew, numerical values approached the following limiting formulas

$$\delta_{n-1,n} = \left(\frac{q\kappa D^2}{(q+1)^2(N+1)}\right)^{\frac{1}{q+1}} n^{-\frac{2}{q+1}}, \qquad \delta_{\star,n} = 0. \tag{4.12}$$

The following theorem proves a convergence guarantee for these choices. Our rates match the lower bounding theory cited in [23] in terms of N, κ, q, D up to a multiplicative constant. Moreover, the coefficient of our convergence rate improves upon prior work [26].

Theorem 4.3. *For any $N \geq 1, D > 0$ and $L(\delta) = \kappa/\delta^q$ and any convex $L(\cdot)$-inexactly smooth function f with minimizer $x_\star$ satisfying $\|x_0 - x_\star\| \leq D$, Algorithm 2 with tolerance sequence δ as in (4.12) is guaranteed to have*

$$f(x_N) - f(x_\star) \leq \frac{\frac{1}{2}D^2 + \sigma_N}{\tau_N} \leq \left(\frac{(q+1)^{\frac{q-1}{q+1}}}{q^{\frac{q}{q+1}}} + o(1)\right) \frac{\kappa^{\frac{1}{q+1}} D^{\frac{2}{q+1}}}{(N+1)^{\frac{2-q}{q+1}}}.$$

where $\sigma_N = \sum_{i=1}^N \tau_{i-1}\delta_{i-1,i}$. In particular, if f is (β, p)-Hölder smooth, Algorithm 2 with suitable

choices of δ has

$$f(x_N) - f(x_\star) \leq \left(\frac{(p+1)^p}{2^{\left(\frac{p+1}{2}\right)}} + o(1) \right) \frac{\beta D^{1+p}}{(N+1)^{\frac{1+3p}{2}}}.$$

The proof of Theorem 4.3, deferred to Appendix C, is a direct generalization of the existing inductive convergence analyses of OGM to include tolerances δ.

Fig. 2 presents two numerical results, fixing $\kappa = D = 1$ for ease. The first shows the PEP value (4.6) for the GFOM and our asymptotic fit of this method. Their worst-case performance is quite similar, even for small N. The second plot shows the effective coefficient of these guarantees for varied q, converging to our asymptotically optimized coefficient as N grows. These motivate the following conjecture on the optimality of our leading coefficient.

Conjecture 4.4. *For any $\kappa, D > 0$, $q \in [0,1]$, $N, d \in \mathbb{N}$ with $d \geq N+2$, and any starting point $x_0 \in \mathbb{R}^d$, there exists a convex, $L(\delta) = \kappa/\delta^q$-inexactly smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $\|x_0 - x_\star\| \leq D$ and a subgradient oracle such that for any subgradient span method,*

$$f(x_N) - f(x_\star) \geq \left(\frac{(q+1)^{\frac{q-1}{q+1}}}{q^{\frac{q}{q+1}}} + o(1) \right) \frac{\kappa^{\frac{1}{q+1}} D^{\frac{2}{q+1}}}{(N+1)^{\frac{2-q}{q+1}}}.$$

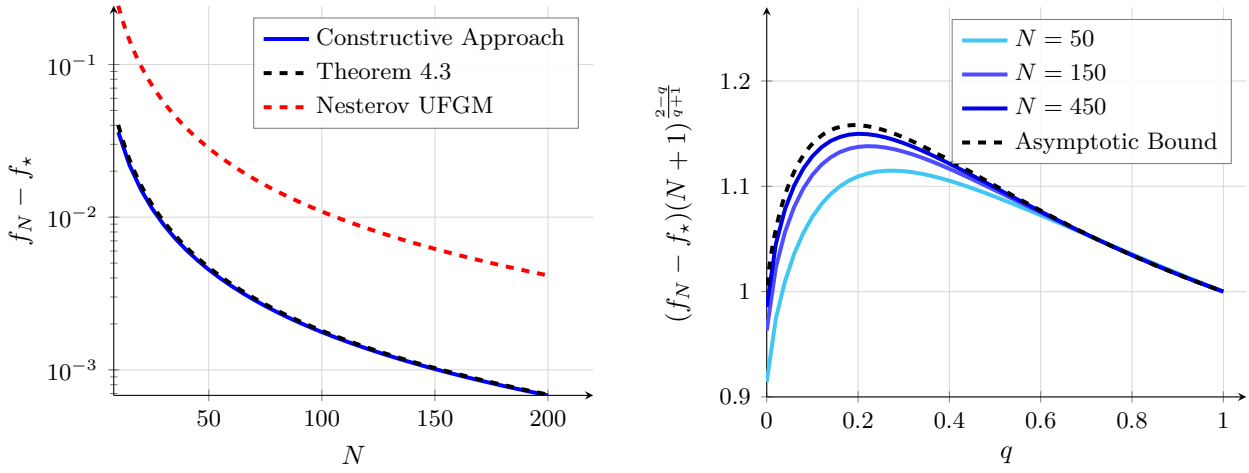

Figure 2: The first plot compares the numerical output of the constructive approach (4.5) for $L(\delta) = 1/\delta^q$ with $q = 0.25$ with Nesterov [26, Theorem 3] and Theorem 4.3. The second displays the conjectured asymptotic tightness of this leading coefficient.

4.4 An Optimized Universal Method for Inexactly Smooth Problems

As a final algorithm, we design an optimized universal and parameter-free method applicable for any $L(\cdot)$ satisfying our assumptions. Universal and parameter-free algorithms offer guarantees for a range of problem instances while requiring no input parameters dependent on their structure (i.e., the function $L(\cdot)$ or optimized tolerances δ). For Hölder smooth convex minimization, such a method (UFGM) was previously designed by [26], generalizing Nesterov's fast gradient method to be universal. Here we similarly extend the OBL method of [27] to handle inexactly smooth functions.

We consider the UOBL method defined in Algorithm 3. The following theorem establishes its convergence guarantee, including an accumulated error term Δ_N identical to [27, Theorem 6], which grows only at each of the logarithmically many backtracking steps.

Algorithm 3 Universal Optimized Backtrackable Linesearch method (UOBL)

Input: x_0, $L_0 > 0$, iteration budget N, and target accuracy $\varepsilon > 0$
Initialize: $\tau_0 = 1/L_0$, $z_1 = x_0 - \tau_0 g_0$
for $n = 1, \ldots, N$
Find the smallest $i \geq 0$ such that, with $L_n = 2^i L_{n-1}$, the values

$$\tau_n = \begin{cases} \tau_{n-1} + \dfrac{1 + \sqrt{1 + 8\tau_{n-1}L_n}}{2L_n} & \text{if } n \leq N-1 \\ \tau_{n-1} + \sqrt{\tau_{n-1}/L_n} & \text{else} \end{cases}$$

$x_n = \frac{\tau_{n-1}}{\tau_n}\left(x_{n-1} - \frac{1}{L_n} g_{n-1}\right) + \frac{\tau_n - \tau_{n-1}}{\tau_n} z_n$
satisfy $f(x_{n-1}) - f(x_n) - \langle g_n, x_{n-1} - x_n \rangle - \frac{1}{2L_n}\|g_{n-1} - g_n\|^2 + \frac{\tau_n - \tau_{n-1}}{\tau_{n-1}} \cdot \frac{\varepsilon}{2} \geq 0$
$z_{n+1} = z_n - (\tau_n - \tau_{n-1}) g_n$

Theorem 4.5. *For any $N \geq 1, D, \varepsilon > 0$ and $L(\cdot)$ satisfying our assumptions, consider a convex $L(\cdot)$-inexactly smooth f with minimizer $x_\star$ satisfying $\|x_0 - x_\star\| \leq D$. Then Algorithm 3 is guaranteed to have*

$$f(x_N) - f(x_\star) \leq \frac{\max\left\{L_0, 2L\left(\frac{\varepsilon}{\sqrt{2}N}\right)\right\}(D^2 + \Delta_N)}{N^2} + \frac{\varepsilon}{2}$$

where $\Delta_N = \sum_{i=1}^N \tau_{i-1}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2$. In particular, if f is $L(\delta) = \kappa/\delta^q$-inexactly smooth, then for L_0 sufficiently small, Algorithm 3 has a big-O optimal convergence rate of

$$f(x_N) - f(x_\star) \leq \frac{2^{1+\frac{q}{2}}\kappa\left(D^2 + \Delta_N\right)}{N^{2-q}\varepsilon^q} + \frac{\varepsilon}{2}.$$

Proof. Consider inexact tolerances $\delta_{i-1,i} := \frac{\tau_i - \tau_{i-1}}{\tau_{i-1}}\frac{\varepsilon}{2}$. The proof of our convergence rate follows inductively, maintaining nonnegativity of the following quantities: for $n = 0, \ldots, N-1$, define

$$H_n = \tau_n(f_\star - f_n + \frac{1}{2L_n}\|g_n\|^2) + \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|z_{n+1} - x_\star\|^2 + \sum_{i=1}^n \tau_{i-1}\delta_{i-1,i} + \sum_{i=1}^n \frac{\tau_{i-1}}{2}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2$$

and for $n = N$, define the modified final value

$$H_N = \tau_N(f_\star - f_N) + \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|z_{N+1} - x_\star\|^2 + \sum_{i=1}^N \tau_{i-1}\delta_{i-1,i} + \sum_{i=1}^N \frac{\tau_{i-1}}{2}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2.$$

With initialization $\tau_0 = \frac{1}{L_0}$ for some $L_0 > 0$ and $z_1 = x_0 - \tau_0 g_0$, the base case holds that $H_0 = \tau_0 \mathcal{C}_{\star,0} \geq 0$ (by convexity of f) with $\mathcal{C}_{\star,0}$ defined below. For $n = 1, \ldots, N$, the induction is maintained by observing the identity

$$H_n = H_{n-1} + \tau_{n-1}\tilde{\mathcal{Q}}_{n-1,n,\delta_{n-1,n}} + (\tau_n - \tau_{n-1})\mathcal{C}_{\star,n}$$

with

$$\tilde{\mathcal{Q}}_{n-1,n,\delta_{n-1,n}} := f_{n-1} - f_n - \langle g_n, x_{n-1} - x_n \rangle - \frac{1}{2L_n}\|g_{n-1} - g_n\|^2 + \frac{\tau_n - \tau_{n-1}}{\tau_{n-1}}\frac{\varepsilon}{2},$$
$$\mathcal{C}_{\star,n} := f_\star - f_n - \langle g_n, x_\star - x_n \rangle.$$

This identity shows that H_n is the sum of three quantities that are nonnegative by definition, and hence H_n is also nonnegative. Then, a convergence rate follows from rearranging $H_N \geq 0$ as

$$f_N - f_\star \leq \frac{\frac{1}{2}D^2 + \sum_{i=1}^N \frac{\tau_{i-1}}{2}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2}{\tau_N} + \frac{\varepsilon}{2} \tag{4.13}$$

by our choices of $\delta_{i-1,i}$ and the nonnegativity of τ_0.

To arrive at our claimed guarantee, we must bound τ_N. For L_0 large enough, $L_n = L_0$ for all $n = 1, \ldots, N$. Therefore we may bound $\tau_N \geq \frac{N^2}{2L_0}$ by considering the recurrence in [27, Corollary 5] for fixed L. Otherwise, suppose we backtrack at least once. Letting $r_n = \frac{\tau_n - \tau_{n-1}}{\tau_{n-1}}$, note that

$$\tau_n - \tau_{n-1} = \begin{cases} \frac{1+\sqrt{1+8\tau_{n-1}L_n}}{2L_n} & n < N \\ \sqrt{\frac{\tau_{n-1}}{L_n}} & n = N \end{cases} \implies r_n \geq \sqrt{\frac{1}{L_n\tau_{n-1}}}, \quad \forall n \leq N.$$

Since the algorithm ensures $\tilde{\mathcal{Q}}_{n-1,n,\delta_{n-1,n}} \geq 0$ for $\delta_{n-1,n} = \frac{(\tau_n - \tau_{n-1})}{\tau_{n-1}}\frac{\varepsilon}{2}$, it follows that

$$L_n \leq 2L\left(\frac{\varepsilon}{2}\frac{(\tau_n - \tau_{n-1})}{\tau_{n-1}}\right) = 2L\left(\frac{\varepsilon}{2}r_n\right) \leq 2L\left(\frac{\varepsilon}{2\sqrt{L_n\tau_{n-1}}}\right) \leq 2L\left(\frac{\varepsilon}{2\sqrt{L_N\tau_N}}\right)$$

where the first inequality holds by the doubling scheme on L_n, the second inequality considers the bound on r_n and the monotonicity of $L(\cdot)$, and the last inequality follows from monotonicity of $L_n\tau_{n-1}$ and τ_N. Therefore, we conclude the uniform bound

$$L_n \leq 2L\left(\frac{\varepsilon}{2\sqrt{L_N\tau_N}}\right), \quad \forall n \leq N.$$

Define as the unique positive solution, $L_\varepsilon(\tau) := \left\{\hat{L} > 0 : \hat{L} = 2L\left(\frac{\varepsilon}{2\sqrt{\hat{L}\tau}}\right)\right\}$, which exists by our monotonicity assumptions on $L(\cdot)$ and convexity of $-1/L(\cdot)$. These same assumptions, along with $L_N \leq 2L\left(\frac{\varepsilon}{2\sqrt{L_N\tau_N}}\right)$ enforce $L_n \leq L_\varepsilon(\tau_N)$ for all $n \leq N$. Therefore

$$\tau_N \geq \frac{N^2}{4L\left(\frac{\varepsilon}{2\sqrt{L_N\tau_N}}\right)} \geq \frac{N^2}{2L_\varepsilon(\tau_N)}$$

by noting the same recurrence as above that τ_N has with fixed L. Finally, taking the unique positive solution to $\hat{\tau}_\varepsilon(N) := \left\{\tau > 0 : \tau = \frac{N^2}{2L_\varepsilon(\tau)}\right\}$, we bound $\tau_N \geq \hat{\tau}_\varepsilon(N)$ from our assumptions on $L(\cdot)$. Using $\hat{\tau}_\varepsilon(N)$ and $L_\varepsilon(\hat{\tau}_\varepsilon(N))$ to simplify (4.13) yields the claimed result.

In the case where f is $L(\delta) = \kappa/\delta^q$-inexactly smooth, we can express these functions exactly. Provided that $L_0 < 2^{1+q/2}\kappa\left(\frac{N}{\varepsilon}\right)^q$, it holds that

$$\begin{aligned} f_N - f_\star &\leq \frac{\frac{1}{2}D^2 + \sum_{i=1}^N \frac{\tau_{i-1}}{2}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2}{\hat{\tau}_\varepsilon(N)} + \frac{\varepsilon}{2} \\ &= \frac{2^{2+\frac{q}{2}}\kappa\left(\frac{1}{2}D^2 + \sum_{i=1}^N \frac{\tau_{i-1}}{2}\left(\frac{1}{L_{i-1}} - \frac{1}{L_i}\right)\|g_{i-1}\|^2\right)}{N^{2-q}\varepsilon^q} + \frac{\varepsilon}{2} \end{aligned}$$

where $L_\varepsilon(\tau) = \left(\frac{2^{q+1}\kappa\tau^{\frac{q}{2}}}{\varepsilon^q}\right)^{\frac{2}{2-q}}$ and $\hat{\tau}_\varepsilon(N) = \frac{N^{2-q}\varepsilon^q}{2^{2+\frac{q}{2}}\kappa}$. □

Acknowledgments. Benjamin Grimmer was supported as an Alfred P. Sloan Foundation fellow.

References

[1] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Cham, 2017.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[3] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8:231–357, 2015.

[4] E. de Klerk, F. Glineur, and A. B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, 2017.

[5] E. de Klerk, F. Glineur, and A. B. Taylor. Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, 30(3):2053–2082, 2020.

[6] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

[7] J. Diakonikolas and C. Guzmán. Optimization on a finer scale: Bounded local subgradient variation perspective. *SIAM Journal on Optimization*, 36:152–184, 2026.

[8] Y. Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

[9] Y. Drori and A. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184:183–220, 2020.

[10] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145:451–482, 2014.

[11] Y. Drori and M. Teboulle. An optimal variant of kelley's cutting-plane method. *Mathematical Programming*, 160:321–351, 2016.

[12] M. P. Friedlander, Goodwin A, and T. Hoheisel. From perspective maps to epigraphical projections. *Mathematics of Operations Research*, 48:1711–1740, 2022.

[13] O. Gannot. A frequency-domain analysis of inexact gradient methods. *Mathematical Programming*, 194:975–1016, 2022.

[14] B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, and A. Dieuleveut. PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python. *Mathematical Programming Computation*, 16:337–367, 2024.

[15] B. Grimmer. On optimal universal first-order methods for minimizing heterogeneous sums. *Optimization Letters*, 18(2):427–445, 2024.

[16] B. Grimmer, K. Shu, and A. L. Wang. Beyond minimax optimality: a subgame perfect gradient method. *Mathematical Programming*, 2026. Published online.

[17] V. Guigues, J. Liang, and R. Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization, 2025.

[18] J. B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer Berlin, Heidelberg, 2001.

[19] T. Hoheisel. Topics in convex analysis in matrix space. Lecture Notes, Spring School on Variational Analysis, Paseky nad Jizerou, Czech Republic, May 2019.

[20] D. Kim and J. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159:81–107, 2016.

[21] T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *Mathematical Programming*, 2025. Published online.

[22] Yin Liu and Sam Davanloo Tajbakhsh. Nonasymptotic analysis of accelerated methods with inexact oracle under absolute error bound, 2025.

[23] A. Nemirovski and Y. Nesterov. Optimal methods of smooth convex minimization. *USSR Comput. Math. Math. Phys.*, 25:21–30, 1985.

[24] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, New York, 1983. Translated from the Russian by E. R. Dawson.

[25] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[26] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152:381–404, 2014.

[27] C. Park and E. Ryu. Optimal first-order algorithms as a function of inequalities. *Journal of Machine Learning Research*, 25:1–66, 2024.

[28] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[29] A. Taylor, J. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27:1283–1313, 2017.

[30] A. Taylor, J. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[31] Y. Wu, Y. Ouyang, Z. Zhang, and Q. Luo. Universal and parameter-free gradient sliding for composite optimization, 2026.

[32] M. Zamani and F. Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35:2182–2201, 2025.

A Deferred Proofs on Inexactly Smooth Calculus

Proof of Lemma 2.4. (1. $\Longrightarrow$ 2.) Here we have that for all x, y with $g_x \in \partial f(x),\ g_y \in \partial f(y)$ and all $\delta \geq 0$

$$f(x) \geq f(y) + \langle g_y, x - y\rangle + \frac{1}{2L(\delta)}\|g_y - g_x\|^2 - \delta.$$

Writing $g_y = g_x + (g_y - g_x)$ and applying Young's inequality to bound the inner product $\langle g_y - g_x, y - x\rangle$ by $\frac{L(\delta)}{2}\|y - x\|^2 + \frac{1}{2L(\delta)}\|g_y - g_x\|^2$, yields the second condition.
(2. $\Longrightarrow$ 3.) It holds that for any x, z and $g_x \in \partial f(x)$ and any $\delta \geq 0$,

$$-f(z) \geq -f(x) + \langle g_x, x - z\rangle - \frac{L(\delta)}{2}\|z - x\|^2 - \delta = f^*(g_x) - \langle z, g_x\rangle - \frac{L(\delta)}{2}\|z - x\|^2 - \delta,$$

where we use the Fenchel-Young equality (2.1). In turn, for any y with $g_y \in \partial f(y)$,

$$\begin{aligned} f^*(g_y) &\geq \langle z, g_y\rangle - f(z) \\ &\geq \langle z, g_y\rangle + f^*(g_x) - \langle z, g_x\rangle - \frac{L(\delta)}{2}\|z - x\|^2 - \delta \\ &= f^*(g_x) + \langle x, g_y - g_x\rangle + \langle z - x, g_y - g_x\rangle - \frac{L(\delta)}{2}\|z - x\|^2 - \delta. \end{aligned}$$

Taking the supremum over z gives $f^*(g_y) \geq f^*(g_x) + \langle x, g_y - g_x \rangle + \frac{1}{2L(\delta)}\|g_y - g_x\|^2 - \delta$.
($3. \implies 1.$) It holds with the Fenchel-Young equality (2.1) that

$$\langle g_y, y \rangle - f(y) = f^*(g_y) \geq \langle g_x, x \rangle - f(x) + \langle x, g_y - g_x \rangle + \frac{1}{2L(\delta)}\|g_x - g_y\|^2 - \delta.$$

Rearranging, swapping the roles of x and y above, and noting this holds for all x, y with respective subgradients g_x, g_y and all $\delta \geq 0$ yields the claimed inexact cocoercive-type condition. □

Proof of Lemma 2.5. We now prove each of the three claims separately.

1. Noting that $g_x \in \partial f(x) \iff \alpha g_x \in \partial(\alpha f)(x)$, applying the change of variables $\delta \mapsto \delta/\alpha$ in (1.4) and rearranging proves the claim.

2. Note that for any $g_{Ax-b} \in \partial f(Ax - b)$, by the subgradient chain rule [1, Corollary 16.72], one has $A^T g_{Ax-b} \in \partial(f \circ (A(\cdot) - b))(x)$. Therefore, with the change of variables $x \mapsto Ax - b$, (1.4) becomes

$$\begin{aligned} f(Ay - b) &\leq f(Ax - b) + \langle g_{Ax-b}, Ay - b - (Ax - b) \rangle + \frac{L(\delta)}{2}\|Ay - b - (Ax - b)\|^2 + \delta \\ &\leq f(Ax - b) + \left\langle A^T g_{Ax-b}, y - x \right\rangle + \frac{\|A\|_{\text{op}}^2 L(\delta)}{2}\|y - x\|^2 + \delta. \end{aligned}$$

3. For any fixed x, let $z_x \in \operatorname{argmin}_z f(x, z)$. Then $F(x) = f(x, z_x)$ and $F(y) = \inf_z f(y, z) \leq f(y, z_x)$. Since $f(x, z)$ is $L(\cdot)$-inexactly smooth it holds for any (x, z), (y, z') and $(g_x, g_z) \in \partial f(x, z)$ and any $\delta \geq 0$ that

$$f(y, z') \leq f(x, z) + \langle g_x, y - x \rangle + \langle g_z, z' - z \rangle + \frac{L(\delta)}{2}(\|x - y\|^2 + \|z - z'\|^2) + \delta.$$

From the subdifferential calculus rule [12, Theorem 1], we know that $\partial F(x) = \{g_x : (g_x, 0) \in \partial f(x, z_x), z_x \in \operatorname{argmin}_z f(x, z)\}$. For any $z_x \in \operatorname{argmin}_z f(x, z)$, and any associated g_x such that $(g_x, 0) \in \partial f(x, z_x)$, let $z = z' = z_x$ and $g_z = 0$. Then, the above inequality establishes the inexact smoothness of F as

$$F(y) \leq f(y, z_x) \leq F(x) + \langle g_x, y - x \rangle + \frac{L(\delta)}{2}\|y - x\|^2 + \delta.$$

□

Proof of Lemma 2.6. Recall from Lemma 2.4 that L_i-inexact smoothness of f_i ensures that all x, y with $g_x^{(i)} \in \partial f_i(x)$ and $g_y^{(i)} \in \partial f_i(y)$ and $\delta_i \geq 0$ have

$$f_i(y) \leq f_i(x) + \left\langle g_x^{(i)}, y - x \right\rangle + \frac{L_i(\delta_i)}{2}\|y - x\|^2 + \delta_i, \tag{A.1}$$

$$f_i^*(g_y^{(i)}) \geq f_i^*(g_x^{(i)}) + \left\langle x, g_y^{(i)} - g_x^{(i)} \right\rangle + \frac{1}{2L_i(\delta_i)}\|g_x^{(i)} - g_y^{(i)}\|^2 - \delta_i. \tag{A.2}$$

1. For any choice of $\delta_i \geq 0$, summing each inequality in (A.1) with $\sum_{i=1}^m \delta_i = \delta$ and utilizing the sum rule [28, Theorem 23.8] establishes a suitable quadratic upper bound for the function $\sum_{i=1}^m f_i$. Taking the infimum over all selections of $\delta_i \geq 0$ gives the claimed formula.

2. Consider the function given by the inner maximum $f^*(g_x) = \max_i f_i^*(g_x)$. By [1, Theorem 18.5], letting $\mathcal{I}^*(g_x) = \{i : f^*(g_x) = f_i^*(g_x)\}$ the subdifferential of this maximum is given by

$$\partial f^*(g_x) = \operatorname{conv} \bigcup_{i \in \mathcal{I}^*(g_x)} \partial f_i^*(g_x).$$

Fix any g_x with $x \in \partial f^*(g_x)$, which the above formula guarantees must take the form $x = \sum_{i \in \mathcal{I}^*(g_x)} \alpha_i x_i$ with $x_i \in \partial f_i^*(g_x)$. Then for any g_y,

$$\begin{aligned} f^*(g_y) \geq \sum_{i \in \mathcal{I}^*(g_x)} \alpha_i f_i^*(g_y) &\geq \sum_{i \in \mathcal{I}^*(g_x)} \alpha_i (f_i^*(g_x) + \langle x_i, g_y - g_x \rangle + \frac{1}{2L_i(\delta)} \|g_y - g_x\|^2 - \delta) \\ &\geq f^*(g_x) + \langle x, g_y - g_x \rangle + \frac{1}{2 \max_i L_i(\delta)} \|g_y - g_x\|^2 - \delta. \end{aligned}$$

Hence f^* is $1/\max_i L_i(\delta)$-inexactly strongly convex and so Lemma 2.4 gives the result.

3. Summing the bounds (A.2) with any selection of $\delta_i \geq 0$ verifies the $\sum 1/L_i(\delta_i)$-inexact strong convexity of $\sum f_i^*$. Taking the supremum over choices of $\delta_i \geq 0$ with $\sum \delta_i = \delta$ tightens this bound, from which the claimed inexact smoothness follows again by Lemma 2.4.

□

Proof of Lemma 2.7. For any $x, y \in \mathbb{R}^d$, $g_x \in \partial f(x)$, and $\delta \geq 0$, it suffices to show that the following S_f function is nonnegative. Namely,

$$0 \leq S_f(x, y, g_x, \delta) := f(x) + \langle g_x, y - x \rangle + \frac{L(\delta)}{2} \|y - x\|^2 + \delta - f(y).$$

Since h is convex with full domain, we can apply a subgradient chain rule [1, Corollary 16.72] to $f = h \circ \|\cdot\|$. From this, any subgradient $g_x \in \partial f(x)$ can be decomposed into the form

$$g_x = \zeta u, \quad \zeta \in \partial h(\|x\|), \quad u \in \partial \|\cdot\|(x).$$

Consider minimizing S_f over all values y with some fixed norm, $\|y\| = s$. Note that if $x \neq 0$ then one has $u = x/\|x\|$ whereas if $x = 0$, one can have any $u \in B(0, 1)$. In either case, since S_f is a simple quadratic with respect to y on this sphere, S_f must minimize over y somewhere collinear with g_x. From this, the result follows as

$$S_f(x, y, g_x, \delta) \geq \min \left\{ h(\|x\|) + \zeta \cdot (\pm\|y\| - \|x\|) + \frac{L(\delta)}{2} (\pm\|y\| - \|x\|)^2 + \delta - h(\|y\|) \right\} \geq 0,$$

where the second inequality holds from $h(|\cdot|)$ being $L(\cdot)$-inexactly smooth. □

B Deferred Proofs for Performance Estimation Theory

Proof of Lemma 3.7. We first show $k = \inf_{\delta \geq 0} \left\{ \frac{L(\delta) + \sqrt{L(\delta)^2 + 2\delta L(\delta)/R^2}}{2} \right\}$ is strictly positive. For $\delta \in [0, 1]$, the term in the infimum is greater than $L(\delta) \geq L(1)$ by monotonicity. For $\delta \in [1, \infty)$, we observe that the expression inside the infimum is also larger than $\frac{\sqrt{\delta L(\delta)}}{R\sqrt{2}}$. To uniformly lower bound this on $[1, \infty)$, notice that the concavity of $1/L(\cdot)$ gives

$$\frac{1}{L(1)} \geq \frac{1}{\delta} \frac{1}{L(\delta)} + \left(1 - \frac{1}{\delta}\right) \frac{1}{L(0)}.$$

Together with $1/L(1) \geq 1/L(0) \geq 0$, where we utilize the convention $1/\infty = 0$, yields

$$\frac{1}{L(\delta)} \leq \frac{1}{L(0)} + \left(\frac{1}{L(1)} - \frac{1}{L(0)}\right)\delta, \qquad \delta \geq 1.$$

Therefore, $\delta L(\delta) \geq \delta / \left(\frac{1}{L(0)} + \left(\frac{1}{L(1)} - \frac{1}{L(0)}\right)\delta\right)$. The right-hand side is increasing in δ, so $\delta L(\delta) \geq L(1)$ for all $\delta \geq 1$. Together, these give the claimed positivity $k \geq \min\left\{L(1), \frac{\sqrt{L(1)}}{R\sqrt{2}}\right\} > 0$.

To verify f is $L(\cdot)$-inexactly smooth, we check the conditions directly. Since the function $h(x)$ is extended linearly outside of the ball with radius R, as a consequence of Lemma 2.7, it suffices to only verify for $\|x\|, \|y\| \leq R$. We note that for given $\|x\| \leq R$, $h(x) = \frac{k}{2}\|x\|^2$ and $\nabla h(x) = kx$. Recall the cocoercive-like condition from Lemma 2.4 for being $L(\cdot)$-inexactly smooth states that for all x, y and $\delta \geq 0$, $h(y) \geq h(x) + \langle \nabla h(x), y - x\rangle + \frac{1}{2L(\delta)}\|\nabla h(x) - \nabla h(y)\|^2 - \delta$. In this case, this simplifies to all $\|x\|, \|y\| \leq R$ and $\delta \geq 0$ having

$$0 \geq \frac{k^2}{2L(\delta)}\|y - x\|^2 - \frac{k}{2}\|y - x\|^2 - \delta.$$

Our choice of $k = \inf_{\delta \geq 0}\left\{\frac{L(\delta) + \sqrt{L(\delta)^2 + 2\delta L(\delta)/R^2}}{2}\right\}$ is exactly the maximum value of k for which this holds. Therefore h is $L(\cdot)$-inexactly smooth. To conclude, we note that so long as $\|Q\|_{\text{op}} \leq 1$, then by Lemma 2.5, $h \circ Q^{1/2}$ is $L(\cdot)$-inexactly smooth as well. □

Proof of Lemma 3.8. We construct a Slater point of a similar form to [30, Theorem 6]. In particular, under the Gram reformulation of (3.7), we show the existence of a strictly feasible point with $G \succ 0$ and $\mathcal{Q}_{i,j} > 0$.

For $L(\cdot)$ satisfying our assumptions consider the construction of an $L(\cdot)$-inexactly smooth function $h(x)$ as defined in Lemma 3.7 for $R = D(1 + \frac{L(1) + \sqrt{L(1)^2 + 2L(1)/D^2}}{2}\|W\|_\infty)^N$. Define

$$Q = \frac{0.5}{2 + 2\cos\left(\frac{\pi}{d+1}\right)}\begin{bmatrix} 2 & 1 & 0 & \dots & 0 \\ 1 & 2 & 1 & \dots & 0 \\ 0 & 1 & 2 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix}, \quad f(x) = h(Q^{1/2}x).$$

By the calculus rules in Lemma 2.5, f is $0.5L(\cdot)$-inexactly smooth and therefore f is $L(\cdot)$-inexactly smooth. Furthermore, f has Hessian kQ for all $\|x\|_Q \leq R$ with $k := \inf_{\delta \geq 0}\left\{\frac{L(\delta) + \sqrt{L(\delta)^2 + 2\delta L(\delta)/R^2}}{2}\right\}$ and minimizer $x_\star = 0$. Choosing $x_0 = De_1$, we show by induction that $\|x_n\|_Q \leq R$ for all $n \leq N$. The base case holds since $\|Q\|_{\text{op}} \leq 1$ and $D \leq R$. Then suppose $\|x_i\|_Q \leq D(1 + \frac{L(1) + \sqrt{L(1)^2 + 2L(1)/D^2}}{2}\|W\|_\infty)^i$ for all $i \leq n - 1$. Consequently,

$$\begin{aligned}
\|x_n\|_Q &\leq \|x_0\|_Q + \|\sum_{i=0}^{n-1} W_{n,i} g_i\|_Q \\
&= D + \|\sum_{i=0}^{n-1} W_{n,i} kQx_i\|_Q \\
&\leq D + k\|W\|_\infty D(1 + k\|W\|_\infty)^{n-1} \\
&\leq D(1 + \frac{L(1) + \sqrt{L(1)^2 + 2L(1)/D^2}}{2}\|W\|_\infty)^n
\end{aligned}$$

where the first inequality uses the triangle inequality, the equality uses $g_i = kQx_i$, the following inequality applies the induction hypothesis and submultiplicativity of the matrix norm, and the last inequality holds by the definition of k, the bound $D \leq R$, and nonnegativity of the norm.

Therefore, $g_n = kQx_n$ for all $n \leq N$. Furthermore, under the assumption $W_{i,i-1} \neq 0$ (see [30, Theorem 6]), it holds that $P = [x_0 \mid g_0 \mid g_1 \mid \ \dots \ \mid g_N]$ is upper triangular with non-zero diagonals and $G = P^T P \succ 0$. Consequently, each gradient is distinct and f belongs to a stricter class than $L(\cdot)$-inexactly smooth functions, having $\mathcal{Q}_{i,j}(F, G) > 0$. With all non-affine constraints strictly feasible, we have constructed a Slater point. □

C Deferred Proofs for Algorithm Design

Proof of the derivation for (4.6)**.** Recall the definition of $p^{\texttt{alg}}_{interp}$ given in (4.5). We can reformulate this by applying the zero equality constraints and noting that $x_\star \in \text{span}\{g_i\}$ implies $\|x_\star\|^2 = \sum_{i=0}^N \frac{\langle g_i, x_\star\rangle^2}{\|g_i\|^2} \leq D^2$. Doing so yields

$$p^{\texttt{alg}}_{interp} = \begin{cases} \max\limits_{f_i, \|g_i\|^2, \langle g_i, x_j\rangle} & f_N - f_\star \\ \text{s.t.} & f_i - f_j - \langle g_j, x_i\rangle - \sup_{\delta_{i,j}\geq 0}\left\{\frac{1}{2L(\delta_{i,j})}\left(\|g_i\|^2 + \|g_j\|^2\right) - \delta_{i,j}\right\} \geq 0, \quad 0 \leq j < i \leq N \\ & f_i - f_j - \sup_{\delta_{i,j}\geq 0}\left\{\frac{1}{2L(\delta_{i,j})}\left(\|g_i\|^2 + \|g_j\|^2\right) - \delta_{i,j}\right\} \geq 0, \quad 0 \leq i < j \leq N \\ & f_\star - f_i - \langle g_i, x_\star\rangle - \sup_{\delta_{\star,i}\geq 0}\left\{\frac{1}{2L(\delta_{\star,i})}\|g_i\|^2 - \delta_{\star,i}\right\} \geq 0, \quad i = 0, \dots, N \\ & f_i - f_\star - \sup_{\delta_{i,\star}\geq 0}\left\{\frac{1}{2L(\delta_{i,\star})}\|g_i\|^2 - \delta_{i,\star}\right\} \geq 0, \quad i = 0, \dots, N \\ & \sum_{i=0}^N \frac{\langle g_i, x_\star\rangle^2}{\|g_i\|^2} \leq D^2. \end{cases} \tag{C.1}$$

The associated dual program is then

$$d^{\texttt{alg}}_{interp} = \begin{cases} \inf\limits_{\delta_{i,j}\geq 0} \min\limits_{\lambda_{i,j}, \lambda_{\star,i}, \lambda_{i,\star}, \nu} \max\limits_{\|g_i\|^2, \langle g_j, x_i\rangle} & \nu D^2 + \sum_{0\leq j<i\leq N} \lambda_{i,j}\left[\delta_{i,j} - \langle g_j, x_i\rangle - \frac{1}{2L(\delta_{i,j})}\left(\|g_i\|^2 + \|g_j\|^2\right)\right] \\ & + \sum_{0\leq i<j\leq N} \lambda_{i,j}\left[\delta_{i,j} - \frac{1}{2L(\delta_{i,j})}\left(\|g_i\|^2 + \|g_j\|^2\right)\right] \\ & + \sum_{i=0}^N \lambda_{\star,i}\left[\delta_{\star,i} - \langle g_i, x_\star\rangle - \frac{1}{2L(\delta_{\star,i})}\|g_i\|^2\right] \\ & + \sum_{i=0}^N \lambda_{i,\star}\left[\delta_{i,\star} - \frac{1}{2L(\delta_{i,\star})}\|g_i\|^2\right] \\ & - \nu \sum_{i=0}^N \frac{\langle g_i, x_\star\rangle^2}{\|g_i\|^2} \\ \text{s.t.} & \sum_{i=j+1}^N \lambda_{j,i} - \sum_{i=0}^{j-1} \lambda_{i,j} = \lambda_{\star,j} - \lambda_{j,\star}, \quad j = 0, \dots, N-1 \\ & \sum_{i=0}^{N-1} \lambda_{i,N} = \sum_{i=0}^{N-1} \lambda_{\star,i} - \lambda_{i,\star} \\ & \sum_{i=0}^N \lambda_{\star,i} - \lambda_{i,\star} = 1 \\ & \lambda_{i,j} \geq 0, \ \lambda_{\star,i} \geq 0, \ \lambda_{i,\star} \geq 0, \ \nu \geq 0. \end{cases} \tag{C.2}$$

where we derive the constraints in the dual program by considering the linear components of the Lagrangian. Moreover, we note that the optimality conditions for the inner maximization imply

that $\lambda_{i,j} = 0$ for $i > j$. Solving the inner maximization yields

$$d_{interp}^{\texttt{alg}} = \begin{cases} \inf\limits_{\delta_{i,j}\geq 0} \min\limits_{\lambda_{i,j},\lambda_{\star,i},\lambda_{i,\star},\nu} & \nu D^2 + \sum_{0\leq i<j\leq N} \lambda_{i,j}\delta_{i,j} + \sum_{i=0}^N \lambda_{\star,i}\delta_{\star,i} + \sum_{i=0}^N \lambda_{i,\star}\delta_{i,\star} \\ \text{s.t.} & \frac{\lambda_{\star,j}^2}{4\nu} \leq \sum_{i=j+1}^N \frac{\lambda_{j,i}}{2L(\delta_{j,i})} + \sum_{i=0}^{j-1} \frac{\lambda_{i,j}}{2L(\delta_{i,j})} + \frac{\lambda_{\star,j}}{2L(\delta_{\star,j})} + \frac{\lambda_{j,\star}}{2L(\delta_{j,\star})}, \quad j = 0,\dots,N \\ & \sum_{i=j+1}^N \lambda_{j,i} - \sum_{i=0}^{j-1} \lambda_{i,j} = \lambda_{\star,j} - \lambda_{j,\star}, \quad j = 0,\dots,N-1 \\ & \sum_{i=0}^{N-1} \lambda_{i,N} = \sum_{i=0}^{N-1} \lambda_{\star,i} - \lambda_{i,\star} \\ & \sum_{i=0}^N \lambda_{\star,i} - \lambda_{i,\star} = 1 \\ & \lambda_{i,j} \geq 0,\ \lambda_{\star,i} \geq 0,\ \lambda_{i,\star} \geq 0,\ \nu \geq 0. \end{cases} \tag{C.3}$$

Substituting $\nu \leftarrow s/2$ and $t_{i,j} \leftarrow \lambda_{i,j}/L(\delta_{i,j})$ gives the claimed dual. □

Proof of Theorem 4.3. Define the following quantity for $n = 0,\dots,N-1$,

$$H_n = \tau_n\left(f_\star - f_n + \frac{1}{2L(\delta_{n,n+1})}\|g_n\|^2\right) + \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|z_{n+1} - x_\star\|^2 + \sum_{i=1}^n \tau_{i-1}\delta_{i-1,i} + \sum_{i=0}^n (\tau_i - \tau_{i-1})\delta_{\star,i}$$

with the following modified value at $n = N$,

$$H_N = \tau_N\left(f_\star - f_N\right) + \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|z_{N+1} - x_\star\|^2 + \sum_{i=1}^N \tau_{i-1}\delta_{i-1,i} + \sum_{i=0}^N (\tau_i - \tau_{i-1})\delta_{\star,i}.$$

We claim that Algorithm 2 inductively maintains the nonnegativity of H_n.

Recall $\mathcal{Q}_{i,j,\delta_{i,j}} := f_i - f_j - \langle g_j, x_i - x_j\rangle - \frac{1}{2L(\delta_{i,j})}\|g_i - g_j\|^2 + \delta_{i,j}$. With $\tau_{-1} := 0$ and for any x_0 and our choice of $\tau_0 = \frac{1}{L(\delta_{0,1})} + \frac{1}{L(\delta_{\star,0})}$, setting $z_1 := x_0 - \tau_0 g_0$, the base case holds that $H_0 = \tau_0 \mathcal{Q}_{\star,0,\delta_{\star,0}} \geq 0$. For any $n = 1,\dots,N$, this nonnegativity is inductively maintained by observing the identity

$$H_n = H_{n-1} + \tau_{n-1}\mathcal{Q}_{n-1,n,\delta_{n-1,n}} + (\tau_n - \tau_{n-1})\mathcal{Q}_{\star,n,\delta_{\star,n}}$$

which shows H_n is a sum of three nonnegative quantities (and hence is nonnegative). This identity follows from the carefully constructed choices of τ_n, x_n, z_n[2].

Hence $H_N \geq 0$. Rearranging the definition of H_N, this establishes a convergence guarantee of

$$f_N - f_\star \leq \frac{\frac{1}{2}\|x_0 - x_\star\|^2 + \sigma_N}{\tau_N} \tag{C.4}$$

where $\sigma_N := \sum_{i=1}^N \tau_{i-1}\delta_{i-1,i} + \sum_{i=0}^N (\tau_i - \tau_{i-1})\delta_{\star,i}$. The remainder of the proof follows from analyzing the sequence τ_n, which is entirely determined by our choice of tolerances $\delta_{i-1,i}$ for $i \leq n$. Below, we carry out these final calculations. Given these, the guarantee outlined for the (β, p)-Hölder smooth setting is a direct result from the $L(\delta) = \kappa/\delta^q$-inexactly smooth guarantee, the implications of Proposition 2.1, and the substitutions $p = \frac{1-q}{1+q}$ and $\beta = \left(\kappa\left(\frac{2}{q}\right)^q\right)^{\frac{1}{(1+q)}}$.

For the chosen, optimized tolerances with

$$\delta_{n-1,n} = \left(\frac{q\kappa D^2}{(q+1)^2(N+1)}\right)^{\frac{1}{q+1}} n^{-\frac{2}{q+1}}, \qquad \delta_{\star,n} = 0. \tag{C.5}$$

[2]As a verification of an effective equivalent identity, see, for example, [16, Lemma 3].

the recurrence defining τ_n collapses to

$$(\tau_n - \tau_{n-1})^2 = \frac{\tau_n}{L(\delta_{n,n+1})} + \frac{\tau_{n-1}}{L(\delta_{n-1,n})} \quad \forall n < N, \quad (\tau_N - \tau_{N-1})^2 = \frac{\tau_{N-1}}{L(\delta_{N-1,N})}. \tag{C.6}$$

Since $\delta_{n-1,n}$ is decreasing in n, $L(\delta_{n-1,n})$ is nondecreasing in n, so for $n < N$,

$$\frac{\tau_n}{L(\delta_{n,n+1})} + \frac{\tau_{n-1}}{L(\delta_{n-1,n})} \geq \frac{\tau_n + \tau_{n-1}}{L(\delta_{n,n+1})} \geq \frac{\left(\sqrt{\tau_n} + \sqrt{\tau_{n-1}}\right)^2}{2L(\delta_{n,n+1})},$$

using $a^2 + b^2 \geq \frac{(a+b)^2}{2}$. Taking the square root in (C.6) and dividing by $\sqrt{\tau_n} + \sqrt{\tau_{n-1}}$ yields

$$\sqrt{\tau_n} - \sqrt{\tau_{n-1}} \geq \frac{1}{\sqrt{2L(\delta_{n,n+1})}}. \tag{C.7}$$

This bound is asymptotically sharp. Bounding (C.6) for $n < N$ from above instead gives

$$\frac{\tau_n}{L(\delta_{n,n+1})} + \frac{\tau_{n-1}}{L(\delta_{n-1,n})} \leq \frac{\tau_n + \tau_{n-1}}{L(\delta_{n-1,n})} \leq \frac{\left(\sqrt{\tau_n} + \sqrt{\tau_{n-1}}\right)^2}{L(\delta_{n-1,n})},$$

using $L(\delta_{n,n+1}) \geq L(\delta_{n-1,n})$ and $\tau_n + \tau_{n-1} \leq (\sqrt{\tau_n} + \sqrt{\tau_{n-1}})^2$. Taking the square root and dividing by $\sqrt{\tau_n} + \sqrt{\tau_{n-1}}$ gives the companion to (C.7): $\sqrt{\tau_n} - \sqrt{\tau_{n-1}} \leq \frac{1}{\sqrt{L(\delta_{n-1,n})}}$ for $n < N$. Substituting (C.5) into $L(\delta) = \kappa/\delta^q$ and dividing by $\sqrt{\tau_{n-1}} \geq \sum_{i=1}^{n-1} \frac{1}{\sqrt{2L(\delta_{i,i+1})}}$ gives

$$0 \leq \frac{\sqrt{\tau_n} - \sqrt{\tau_{n-1}}}{\sqrt{\tau_{n-1}}} \leq \frac{\sqrt{2} n^{-\frac{q}{q+1}}}{\sum_{i=1}^{n-1} (i+1)^{-\frac{q}{q+1}}} = \frac{\sqrt{2}}{q+1} n^{-1} \left(1 + o(1)\right) \to 0,$$

where the equality uses $\sum_{i=1}^{n} (i+1)^{-\frac{q}{q+1}} = (q+1)(n+1)^{\frac{1}{q+1}}(1 + o(1))$. Therefore, $\tau_n/\tau_{n-1} \to 1$ uniformly in N. Since $L(\delta_{n,n+1})/L(\delta_{n-1,n}) = (1 + 1/n)^{\frac{2q}{q+1}} \to 1$ as well, (C.6) reads $(\tau_n - \tau_{n-1})^2 = \frac{2\tau_n}{L(\delta_{n,n+1})}(1 + o(1))$, and with $\sqrt{\tau_n} + \sqrt{\tau_{n-1}} = 2\sqrt{\tau_n}(1 + o(1))$ this yields the asymptotic equality $\sqrt{\tau_n} - \sqrt{\tau_{n-1}} = \frac{1}{\sqrt{2L(\delta_{n,n+1})}}(1 + o(1))$ for $n < N$. Telescoping these terms gives

$$\tau_n = \frac{(q+1)^{\frac{2}{q+1}} \left(qD^2\right)^{\frac{q}{q+1}}}{2\kappa^{\frac{1}{q+1}}} \frac{(n+1)^{\frac{2}{q+1}}}{(N+1)^{\frac{q}{q+1}}} \left(1 + o(1)\right), \quad \forall n \leq N, \tag{C.8}$$

where the modified case of $n = N$ only adds lower order terms. Considering our choices of $\delta_{n-1,n}$, it holds $\tau_{n-1}\delta_{n-1,n} = \frac{qD^2}{2(N+1)}(1 + o(1))$ for each $n \leq N$ so

$$\sigma_N = \sum_{i=1}^{N} \tau_{i-1}\delta_{i-1,i} = \frac{qD^2}{2} \left(1 + o(1)\right). \tag{C.9}$$

Substituting (C.8) and (C.9) at $n = N$ into (C.4) gives the claimed result. □