Frank–Wolfe Beyond $1/t$ Convergence

Sebastian Pokutta
Institute of Mathematics, Technische Universität Berlin and
Zuse Institute Berlin, Germany
pokutta@zib.de

Abstract

We study the vanilla Frank–Wolfe algorithm for smooth convex minimization over compact convex sets. Well-known lower bounds establish a uniform worst-case $\Omega(1/t)$ primal-gap barrier in the general smooth convex case, and faster convergence usually requires favorable function properties such as Hölder error bounds or strong convexity. We present a new *Local Dual Sharpness (LDS)* condition on the feasible region and the LMO used by the algorithm. For every fixed smooth convex instance satisfying LDS, Frank–Wolfe satisfies $f(x_t)-f^\star = o(1/t)$ under short steps, exact line search, and the classical open-loop rule. This excludes an asymptotic $\Omega(1/t)$ trajectory lower bound on any one such instance; this is not to be confused with a uniform minimax statement. The condition generalizes and localizes uniform convexity of sets and is satisfied globally by every uniformly convex set. To our knowledge, this gives the first fixed-instance $o(1/t)$ result on uniformly convex sets without an objective growth, gradient-lower-bound, or strong-convexity assumption. Combining LDS with a local Hölder error bound quantifies the rate.

1 Introduction

We consider optimization problems of the form $\min_{x\in\mathcal{C}} f(x)$, where f is smooth and convex and $\mathcal{C}$ is a compact convex set. The Frank–Wolfe algorithm [Frank and Wolfe, 1956], also called the conditional gradient method [Levitin and Polyak, 1966], is one of the classical projection-free methods for smooth constrained convex optimization; see Braun et al. [2025], Pokutta [2024] for an overview. Its generic convergence rate on a compact convex set is $O(1/t)$, where the lower-bound instance arises from minimizing $f(x) = \|x\|_2^2$ over the probability simplex [Jaggi, 2013, Lan, 2013]. Higher convergence rates typically require assumptions on the structure of f, e.g., strong convexity or lower-bounded gradients, but it is *a priori* not clear which assumptions are necessary. For example, only recently in Halbey et al. [2026] it was shown that in the case of strongly convex functions over smooth strongly convex sets the worst-case rate can be as bad as $\Omega(1/t^2)$, answering a long-standing open question arising from Garber and Hazan [2015]; shortly afterwards it was shown that for nonsmooth strongly convex sets this even holds for any LMO-based method, i.e., those that access the feasible region solely through linear minimization [Grimmer and Liu, 2026]. The overall convergence landscape is complex for the Frank–Wolfe algorithm and we discuss it in detail in the related work section.

In this work we introduce the *Local Dual Sharpness (LDS)* condition, imposed on the feasible region together with its Linear Minimization Oracle (LMO) around a nonempty reference set M. The definition and our conditional theorems allow a general M; in the principal instance-level results we take $M = \mathcal{M} = \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Thus, for locally curved sets, applicability depends on where the objective places its minimizers; we discuss the precise assumptions below. Once LDS holds around $\mathcal{M}$, Frank–Wolfe with exact line search, short steps, or the classical open-loop step-size satisfies $f(x_t) - f^\star = o(1/t)$ for that fixed smooth convex instance, without any additional objective growth,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Regime	Representative rate picture
Compact-convex baseline	Generic $O(1/t)$ convergence, with matching $\Omega(1/t)$ lower bounds showing that no uniform improvement is possible without additional structure.
Location-based acceleration	Interior minimizers yield linear short-step / line-search rates. On UC(q) sets with lower-bounded gradient, short step, line search, and open loop become linear or faster polynomial.
Curved sets plus growth	HEB together with UC(q) yields explicit polynomial improvements. In the SC / SC case this recovers $O(1/t^2)$, and both small-scale and large-scale lower bounds are now known.
Proper-face regime	On polytopes, short step and exact line search can be nearly $1/t$, whereas open loop can still accelerate after active-face identification.
This paper: LDS	For every fixed instance satisfying LDS around its minimizer set, the primal gap is $o(1/t)$. LDS(q) plus a local HEB yields explicit polynomial tails.

Table 1: Compressed overview of the Frank–Wolfe rate landscape most relevant here. Section A.1 and Table 2 give the full row-by-row comparison and source discussion.

strong-convexity, or gradient-lower-bound assumption. For globally uniformly convex sets LDS holds around every reference set, and hence the conclusion applies to every smooth convex objective. The improvement is qualitative and instance-dependent. In particular, we do not claim a uniform burn-in or minimax modulus. Quantitative rates are recovered below by adding a local Hölder error bound.

Related Work and State of the Art

The classical compact-convex story is by now well understood: vanilla Frank–Wolfe has an $O(1/t)$ upper bound, and matching $\Omega(1/t)$ lower bounds follow from simplex-type constructions and deterministic first-order oracle arguments [Frank and Wolfe, 1956, Jaggi, 2013, Lan, 2013, Levitin and Polyak, 1966]. Faster rates traditionally require additional structure. The most classical examples are interior minimizers and the lower-bounded-gradient regime on curved sets, which lead to linear or accelerated polynomial rates under short step, exact line search, and open loop [Garber and Hazan, 2015, Guélat and Marcotte, 1986, Kerdreux et al., 2021, Wirth et al., 2023, 2025, Wolfe, 1970].

A complementary body of work shows that curvature of the feasible region alone does not settle the picture. Uniformly convex sets combined with growth or error-bound assumptions yield explicit polynomial improvements [Kerdreux et al., 2021, Peña, 2023], while recent uniform finite-horizon lower bounds show that even strongly convex objectives over strongly convex sets can be as slow as $\Omega(1/t^2)$ in the worst case [Grimmer and Liu, 2026, Halbey et al., 2026]. The interior-minimizer and lower-bounded-gradient linear regimes do not cover the boundary zero-gradient case used by these hard instances. In another direction, proper-face phenomena on polytopes can slow short step and exact line search to nearly $1/t$, whereas open-loop rules can still exhibit accelerated convergence once the active face has been identified [Canon and Cullum, 1968, Wirth et al., 2025, Wolfe, 1970]. The construction in Zimmer et al. [2026] provides a complementary low-dimensional worst-case instance for uniformly convex sets and quadratic objectives.

A comprehensive overview and comparison had to be relegated to Section A.1.

Contribution

Once LDS is assumed around the objective-dependent minimizer set, the convergence result requires no additional *growth or curvature* condition on f beyond smoothness and convexity. For uniformly convex feasible sets, which satisfy LDS around every reference set, this qualification is automatic.

Local Dual Sharpness. We introduce the *Local Dual Sharpness (LDS)* condition, which subsumes uniform convexity of sets and also applies to suitable locally curved regions. For every fixed smooth convex instance satisfying LDS around its minimizer set, Frank–Wolfe converges as $o(1/t)$ without an objective growth condition. Equivalently, no such fixed instance has a persistent asymptotic $\Omega(1/t)$ primal-gap trajectory. This does not imply that $t \sup_I F_t(I) \to 0$ (with $F_t(I)$ denoting the primal

gap of instance I after t steps) over a class whose LDS constants, objective, geometry, or burn-in vary with I, i.e., it is not a minimax statement. To the best of our knowledge, the uniformly-convex specialization is the first fixed-instance $o(1/t)$ result for every smooth convex objective on such a set without a Hölder error bound, strong convexity, or a lower-bounded gradient. Prior accelerated results over uniformly convex sets assumed one of these additional objective conditions.

Our local condition should be contrasted with the local scaling perspective in [Dunn, 1979, Kerdreux et al., 2026], where the locality is attached to a boundary point of $\mathcal{C}$ and a specific normal direction and generally no acceleration is obtained without additional assumptions on the function. LDS is instead imposed around an entire reference set and uniformly over all LMO directions; when the LMO minimizer is nonunique, its atom-selection rule is fixed as specified below.

Robustness under Step-size Choices. We prove our results for the three common step-size strategies: short steps arising from optimizing over the smoothness inequality and (exact) line search in Corollary 3.8; for open-loop step-sizes, Theorem 3.10 considers the classical open-loop $\gamma_t = 2/(t+2)$ and Remark A.7 treats the generalized family $\gamma_t = \ell/(t+\ell)$. The open-loop proof is significantly more involved, since Frank–Wolfe need not be a descent method in that case.

Quantitative Rates. Finally we combine LDS with a downstream local Hölder error bound to quantify our $o(1/t)$ rates, yielding new accelerated convergence regimes for Frank–Wolfe. While the obtained accelerated regimes are broader than previously known ones, this quantification comes at a cost of strengthened assumptions for f. The obtained quantitative bounds are tail estimates as customary (see [Braun et al., 2025]): after a finite burn-in t_0 the stated rate holds for all $t \geq t_0$, and we do not optimize this entry time. In particular, we obtain accelerated sublinear rates of the form $O(t^{-rq/(rq-2)})$, where r is the local Hölder error bound order and q is the LDS power. In Section A.4 we complement these quantitative rates by showing that $o(1/t)$ rates obtained via LDS can be slower than any polynomial rate of form $O(t^{-1-\eta})$ for any $\eta > 0$.

We formally verified the principal arguments in Lean 4 (see e.g., [Moura and Ullrich, 2021]). A [LEAN] badge denotes either an exact named endpoint or an explicitly identified collection of verified ingredients supporting the displayed statement. For Corollary 3.8, the zero-gradient and lower-bounded-gradient branches are formalized separately and combined by the elementary compactness argument shown in the paper. We do not formalize either the standard preprocessing step of passing to orthonormal coordinates on $\operatorname{aff}(\mathcal{C})$ or the concrete geometry of the illustrative sets. The verification will be made available on GitHub, and reusable components will be proposed upstream to mathlib. Some proofs are in the appendix due to space; the Lean files are supplementary material.

2 Preliminaries

We briefly summarize the notions that we will use throughout and we refer the reader to Braun et al. [2025] for an in-depth treatment. We consider the constrained convex optimization problem $\min_{x\in\mathcal{C}} f(x)$, where $\mathcal{C} \subseteq \mathbb{R}^d$ is compact and convex and $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and continuously differentiable. For exposition we work in coordinates on the affine hull of $\mathcal{C}$; equivalently, the definitions below may be read relative to $\operatorname{aff}(\mathcal{C})$, with directions and gradient norms projected to its linear span. We define $f^\star \stackrel{\text{def}}{=} \min_{x\in\mathcal{C}} f(x)$, $\mathcal{M} \stackrel{\text{def}}{=} \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and $D \stackrel{\text{def}}{=} \max_{x,y\in\mathcal{C}} \|x-y\|$, where $f^\star$ is the *optimal objective value*, $\mathcal{M}$ is the *set of minimizers*, D is the *diameter* of $\mathcal{C}$, and $\|\cdot\|$ is the Euclidean norm. Throughout, we assume that f is L-smooth on $\mathcal{C}$ for some $L > 0$, i.e.,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2 \qquad \forall x, y \in \mathcal{C}. \tag{2.1}$$

Given an iterate $x_t \in \mathcal{C}$, let $s_t \in \operatorname{argmin}_{s\in\mathcal{C}} \langle \nabla f(x_t), s\rangle$ be a *Frank–Wolfe atom* (also called *Frank–Wolfe vertex* in the case of polytopes) that is returned by a *Linear Minimization Oracle (LMO)* for $\mathcal{C}$. If the argmin set is not a singleton, we fix a *deterministic single-valued selector* $S(g) \in \operatorname{argmin}_{s\in\mathcal{C}} \langle g, s\rangle$ and set $s_t = S(\nabla f(x_t))$. Note that the LDS property below is defined with respect to a given fixed selector S; in particular LDS for one tie-breaking rule does not necessarily imply LDS for another rule. We write $F_t \stackrel{\text{def}}{=} f(x_t) - f^\star = f(x_t) - f(x^\star)$ as the *primal gap*, where $x^\star$ is any optimal solution, $g_t \stackrel{\text{def}}{=} \langle \nabla f(x_t), x_t - s_t\rangle$ as the *Frank–Wolfe gap*, and $d_t \stackrel{\text{def}}{=} \|x_t - s_t\|$. By convexity we immediately obtain $0 \leq F_t \leq g_t$.

We will be concerned with the Frank–Wolfe algorithm [Frank and Wolfe, 1956, Levitin and Polyak, 1966], stated in Algorithm 1. The key point is that it is *projection-free* by forming its updates as convex combinations of the form $x_{t+1} = x_t + \gamma_t(s_t - x_t)$.

For the choice of the step-size rule in Line 3 of Algorithm 1, we will consider the following three rules:

$$\gamma_t^{\mathrm{ss}} \stackrel{\mathrm{def}}{=} \min\left\{\frac{g_t}{Ld_t^2}, 1\right\}, \qquad \textit{(short steps)} \tag{2.2}$$

$$\gamma_t^{\mathrm{ls}} \in \operatorname*{argmin}_{\gamma\in[0,1]} f\big(x_t + \gamma(s_t - x_t)\big), \qquad \textit{(exact line search)} \tag{2.3}$$

$$\gamma_t^{\mathrm{ol}} \stackrel{\mathrm{def}}{=} \frac{2}{t+2}. \qquad \textit{(open-loop)} \tag{2.4}$$

We will later also consider generalized open-loop step-sizes of the form $\gamma_t \stackrel{\mathrm{def}}{=} \frac{\ell}{t+\ell}$ with $\ell \in \mathbb{N}$.

The following properties and results are classical; we refer the reader to Braun et al. [2025].

Algorithm 1 The Frank–Wolfe algorithm [Frank and Wolfe, 1956]

Input: $x_0 \in \mathcal{C}$ and a rule for choosing $\gamma_t \in [0,1]$

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: $\quad s_t \leftarrow \operatorname{argmin}_{s\in\mathcal{C}} \langle \nabla f(x_t), s\rangle$
3: $\quad$ choose γ_t according to (2.2), (2.3), or (2.4)
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t(s_t - x_t)$
5: **end for**

Proposition 2.1 (Primal progress estimate [LEAN])**.** *Let $x_{t+1} = x_t + \gamma_t(s_t - x_t)$ with $\gamma_t \in [0,1]$. Then $F_{t+1} \le F_t - \gamma_t g_t + \frac{L}{2}\gamma_t^2 d_t^2$. In particular:*

1. *for the short-step rule and exact line search, we have $F_{t+1} \le F_t - \frac{1}{2}\min\left\{g_t, \frac{g_t^2}{Ld_t^2}\right\}$;*
2. *for the open-loop rule, we have $F_{t+1} \le \left(1 - \frac{2}{t+2}\right) F_t + \frac{L}{2}\left(\frac{2}{t+2}\right)^2 d_t^2$.*

The next proposition states the standard primal convergence guarantee for the Frank–Wolfe algorithm due to Frank and Wolfe [1956], Levitin and Polyak [1966], from the first iterate onward. The proof is standard; see e.g., [Braun et al., 2025, Jaggi, 2013, Pokutta, 2024].

Proposition 2.2 ($O(1/t)$ primal convergence for the Frank–Wolfe algorithm [LEAN])**.** *Let f be an L-smooth and convex function for some $L > 0$, and let $\mathcal{C}$ be a non-empty compact convex set. Running Algorithm 1, for each of the three step-size rules (2.2), (2.3), and (2.4), the iterates x_t satisfy*

$$F_t \le \frac{2LD^2}{t+2} \qquad \forall t \ge 1; \qquad \textit{in particular } F_t \to 0. \tag{2.5}$$

We will also need the next simple observation to deal with possibly non-unique minimizers.

Lemma 2.3 (Distance to the minimizer set [LEAN])**.** *Under the assumptions from Proposition 2.2, further assume $F_t \to 0$ and let $\mathcal{M} = \{x \in \mathcal{C} : f(x) = f^\star\}$ be the set of minimizers. Then $\operatorname{dist}(x_t, \mathcal{M}) \to 0$.*

Finally, we will be concerned with uniformly convex sets as feasible regions.

Definition 2.4 (Power-type uniform convexity)**.** Let $q \in \mathbb{R}$ satisfy $q \ge 2$, and let $\alpha > 0$. We say that $\mathcal{C}$ is (α, q)-*uniformly convex* if for every $x, y \in \mathcal{C}$, every $\lambda \in [0,1]$, and every $z \in \mathbb{R}^d$ with $\|z\| = 1$,

$$\lambda x + (1-\lambda)y + \lambda(1-\lambda)\alpha\,\|x - y\|^q\, z \in \mathcal{C}. \tag{2.6}$$

The case $q = 2$ is exactly the usual *strong convexity* of the feasible set. Thus every strongly convex set is uniformly convex, while the definition also includes genuinely higher-order curved bodies such as ℓ_p balls for $p > 2$ [Kerdreux et al., 2021]. The next lemma turns contractions into rates.

Lemma 2.5 (Discrete power-law descent [LEAN]). *Let (a_t) be a nonnegative sequence. Assume that for some $0 < \eta \leq 1$, $0 < r \leq 1$, and some $t_0 \in \mathbb{N}$ with $a_{t_0} \leq 1$, it holds $a_{t+1} \leq a_t - \eta a_t^{1+r}$ for all $t \geq t_0$. Then $a_t = O\left(t^{-1/r}\right)$.*

Proof deferred to Section A.7 (Deferred Proofs).

3 Local Dual Sharpness

The following definition introduces the *Local Dual Sharpness (LDS)* condition that we will be working with; we write M for a general set and $\mathcal{M}$ for the set of minimizers.

Definition 3.1 (Local dual sharpness around a reference set M). Let $M \subseteq \mathcal{C}$ be nonempty, let $A > 0$, and let $q \in \mathbb{R}$ satisfy $q \geq 2$. We say that $\mathcal{C}$ together with its fixed LMO atom-selection rule satisfies *local dual sharpness around M with constants (A, q)* if there exists $\rho > 0$ such that whenever $x \in \mathcal{C}$ with $\operatorname{dist}(x, M) < \rho$, then for every $g \in \mathbb{R}^d$ and the atom $s \in \operatorname{argmin}_{y \in \mathcal{C}} \langle g, y \rangle$ returned by that fixed LMO selection, it holds:

$$A \|g\| \, \|x - s\|^q \leq \langle g, x - s \rangle .$$

Remark 3.2 (Dependence on the atom-selection rule). Local dual sharpness is a property of the support map induced by the atom-selection rule of the LMO, not just of the feasible region $\mathcal{C}$. Indeed, for fixed x and g, the Frank–Wolfe dual gap $\langle g, x - s \rangle = \langle g, x \rangle - \min_{y \in \mathcal{C}} \langle g, y \rangle$ is the same for every minimizing atom $s \in \operatorname{argmin}_{y \in \mathcal{C}} \langle g, y \rangle$. What can change across selections though is the displacement $\|x - s\|$. The subtlety disappears when the supporting atom is unique, but if the exposed face is flat then different selections on that face can produce different displacements and therefore different admissible constants (A, q), or even validity for one selection and failure for another. We thus assume that the selection rule is deterministic and fixed to avoid this complication.

Support-function form. Let $S(g)$ be the fixed selector and let $\sigma_{\mathcal{C}}(u) \stackrel{\text{def}}{=} \sup_{y \in \mathcal{C}} \langle u, y \rangle$ be the support function. Then LDS is equivalently

$$\langle g, x \rangle + \sigma_{\mathcal{C}}(-g) \geq A \|g\| \, \|x - S(g)\|^q$$

near M. If the supporting point is unique, then $S(g) = \nabla \sigma_{\mathcal{C}}(-g)$ and this is a power-type exposure inequality for the support function.

Checking whether LDS holds means certifying the LDS inequality simultaneously for every x in a neighborhood of M and every normalized direction g. This paper does not study the computational complexity of that certification problem. Useful *sufficient certificates* include global uniform convexity; a uniform affine-hull interior-ball condition around M; the curved-patch plus residual-gap criterion in Proposition A.2; or a direct positive lower bound on the normalized ratio over $\|g\| = 1$. If $M = \mathcal{M}$ is not known, it must first be localized by problem-specific optimality information. Note, however, that the Frank–Wolfe algorithm is adaptive, i.e., it automatically benefits if LDS holds, not requiring certification.

Remark 3.3 (Beyond lower-bounded gradients $\nabla f(\mathcal{C}) > 0$). Assume that LDS holds around the minimizer set $\mathcal{M}$ with constants (A, q). If the gradients were bounded away from zero, i.e., $\|\nabla f(x_t)\| \geq G > 0$, then once x_t is close enough to $\mathcal{M}$ the LDS inequality would give $AGd_t^q \leq g_t$, so that $g_t \to 0$ would force $d_t \to 0$. This would recover the accelerated convergence of [Demyanov and Rubinov, 1970, Levitin and Polyak, 1966]. Here, however, we consider the more involved general case without such assumptions, so that $\|\nabla f(x_t)\|$ may itself tend to zero near $\mathcal{M}$.

Proposition 3.4 (Uniform convexity implies local dual sharpness [LEAN]). *If $\mathcal{C}$ is (α, q)-uniformly convex, then for every nonempty $M \subseteq \mathcal{C}$, any LMO with any fixed atom-selection rule satisfies local dual sharpness around M with constants $(\alpha/2, q)$.*

Proof. By Lemma A.1, $\frac{\alpha}{2} \|g\| \, \|x - s\|^q \leq \langle g, x - s \rangle$ holds for every $x \in \mathcal{C}$, every $g \in \mathbb{R}^d$, and every selected minimizing atom s. Thus the inequality holds globally for any $\rho > 0$ in Definition 3.1. □

The local dual sharpness condition is more general than (global) uniform convexity; we present two examples in Section A.3 in the appendix. Its validity around $\mathcal{M}$, however, depends on the location of that objective-dependent set, e.g., in Figure 1 it holds on the rounded caps but not on a flat side. Strict

convexity alone also does not imply LDS; see Remark A.4. Our setup differs from the local uniform convexity and local scaling perspective of [Dunn, 1979, Kerdreux et al., 2026], where locality is expressed at one boundary point and a singled-out normal direction. Here the condition is imposed on the fixed LMO support map around an entire reference set and over all LMO directions.

3.1 Short Steps and Exact Line Search

Here and in the following, we often have to pay special attention to the potential $|\mathcal{M}| > 1$ case. We make this explicit by splitting theorems.

Theorem 3.5 (LDS + short steps or exact line search / conditional form [LEAN]). *Let $M \subseteq \mathcal{C}$ be nonempty. Assume that the Frank–Wolfe iterates are generated either by the global short-step rule or by exact line search. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around M with some constants (A, q). Let $x^\star \in \mathcal{C}$, and let $(p_t) \subseteq M$ satisfy $f(p_t) = f(x^\star)$ and $\nabla f(p_t) = 0$ for all t. If $\delta_t \overset{\text{def}}{=} \|x_t - p_t\| \to 0$, then*

$$(t+2)\big(f(x_t) - f(x^\star)\big) \to 0.$$

Proof. Define $H_t \overset{\text{def}}{=} \frac{1}{F_t}$, where F_t is the primal gap. If $F_t = 0$ for some t, then the claim is immediate. So we assume $F_t > 0$ for all t. Let $\rho > 0$ be given by Definition 3.1. Since $\operatorname{dist}(x_t, M) \le \|x_t - p_t\| = \delta_t \to 0$, there exists t_0 such that $\operatorname{dist}(x_t, M) < \rho$ for all $t \ge t_0$. Hence for every $t \ge t_0$, local dual sharpness yields $A \|\nabla f(x_t)\| \, d_t^q \le g_t$.

Convexity at the zero-gradient minimizer p_t gives $F_t = f(x_t) - f(p_t) \le \langle \nabla f(x_t), x_t - p_t \rangle \le \|\nabla f(x_t)\| \, \delta_t$. Moreover, since $\nabla f(p_t) = 0$ and f is L-smooth, we have $F_t \le \frac{L}{2}\delta_t^2$. Combining the first estimate with the local dual sharpness inequality gives

$$A \frac{F_t}{\delta_t} d_t^q \le g_t \qquad \forall t \ge t_0. \tag{3.1}$$

By Proposition 2.1, the short-step and exact-line-search rules both satisfy

$$F_{t+1} \le F_t - \frac{1}{2} \min\left\{ g_t, \frac{g_t^2}{L d_t^2} \right\}.$$

Therefore

$$H_{t+1} - H_t = \frac{F_t - F_{t+1}}{F_t \, F_{t+1}} \ge \frac{F_t - F_{t+1}}{F_t^2},$$

where the latter inequality holds because F_t is monotone decreasing. We split into two cases.

Case 1: $g_t \ge L d_t^2$. Then $F_t - F_{t+1} \ge \frac{1}{2} g_t \ge \frac{1}{2} F_t$, so $H_{t+1} - H_t \ge \frac{1}{2F_t} \ge \frac{1}{L\delta_t^2}$ follows.

Case 2: $g_t \le L d_t^2$. Then $F_t - F_{t+1} \ge \frac{g_t^2}{2 L d_t^2}$. Using $g_t \ge F_t$ and (3.1) we obtain the two inequalities

$$H_{t+1} - H_t \ge \frac{g_t^2}{2 L d_t^2 F_t^2} \ge \frac{1}{2 L d_t^2}, \qquad \text{and} \qquad H_{t+1} - H_t \ge \frac{A^2 d_t^{2q-2}}{2 L \delta_t^2}. \tag{3.2}$$

Let $\theta = (q-1)/q$ and since $\max\{u, v\} \ge u^\theta v^{1-\theta}$ for all $u, v \ge 0$, (3.2) implies

$$H_{t+1} - H_t \ge \left(\frac{1}{2 L d_t^2} \right)^{(q-1)/q} \left(\frac{A^2 d_t^{2q-2}}{2 L \delta_t^2} \right)^{1/q} = \frac{A^{2/q}}{2L} \, \delta_t^{-2/q}.$$

After increasing t_0 if needed, we may also assume $\delta_t < 1$ for all $t \ge t_0$. Since $q \ge 2$, this gives $\delta_t^{-2} \ge \delta_t^{-2/q}$, so Case 1 also yields $H_{t+1} - H_t \ge \frac{1}{L} \, \delta_t^{-2/q}$. Thus both cases imply $H_{t+1} - H_t \ge \min\left\{ \frac{1}{L}, \frac{A^{2/q}}{2L} \right\} \delta_t^{-2/q}$ for all $t \ge t_0$. Since $\delta_t \to 0$, the reciprocal increments $H_{k+1} - H_k$ diverge to $+\infty$ along the tail. Their Cesàro averages therefore diverge as well, i.e.,

$$\frac{H_t}{t} - \frac{H_0}{t} = \frac{1}{t} \sum_{k=0}^{t-1} (H_{k+1} - H_k) \to +\infty.$$

Thus $\frac{H_t}{t} \to \infty$, equivalently $t F_t \to 0$. In particular $H_t \to \infty$, hence $F_t \to 0$. Therefore $(t+2) F_t = t F_t + 2 F_t \to 0$. □

Theorem 3.6 (LDS + short steps or line search with a zero-gradient minimizer [LEAN]). *Let $\mathcal{M}$ be the minimizer set. Assume that the Frank–Wolfe iterates are generated either by the global short-step rule or by exact line search. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around $\mathcal{M}$ with some constants (A, q). If $\mathcal{M}$ contains a point $x^\star$ with $\nabla f(x^\star) = 0$, then $f(x_t) - f^\star = o(1/t)$.*

Proof. By Proposition 2.2, $F_t \to 0$. Hence Lemma 2.3 gives $\operatorname{dist}(x_t, \mathcal{M}) \to 0$. Choose $p_t \in \mathcal{M}$ with $\|x_t - p_t\| = \operatorname{dist}(x_t, \mathcal{M})$. Then $\delta_t \stackrel{\text{def}}{=} \|x_t - p_t\| \to 0$. Since $x^\star$ is a zero-gradient point of a differentiable convex function, it is a global minimizer on the ambient space. Therefore every point of $\mathcal{M}$ is also a zero-gradient minimizer, so $\nabla f(p_t) = 0$ for all t. Applying Theorem 3.5 yields $(t+2)\big(f(x_t) - f^\star\big) \to 0$. □

The proof of the next proposition is similar to Demyanov and Rubinov [1970], Garber and Hazan [2015], Levitin and Polyak [1966]; see also Braun et al. [2025].

Proposition 3.7 (LDS + short steps or exact line search when $\nabla f(\mathcal{C}) > 0$ [LEAN]). *Let $\mathcal{M}$ be the minimizer set. Assume that the Frank–Wolfe iterates are generated either by the global short-step rule or by exact line search. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around $\mathcal{M}$ with some constants (A, q) and that $G \stackrel{\text{def}}{=} \min_{x \in \mathcal{C}} \|\nabla f(x)\| > 0$. Then:*

1. *if $q = 2$, then with $\kappa \stackrel{\text{def}}{=} \frac{1}{2} \min\left\{1, \frac{AG}{L}\right\}$, there exists $t_0 \in \mathbb{N}$ such that $F_t \le (1-\kappa)^{t-t_0} F_{t_0}$ for all $t \ge t_0$.*
2. *if $q > 2$, then there exist $t_0 \in \mathbb{N}$ and $K > 0$ such that $F_t \le \frac{K}{(t-t_0+1)^{q/(q-2)}}$ for all $t \ge t_0$.*

Proof deferred to Section A.7 (Deferred Proofs).

Combining Theorem 3.6 and Proposition 3.7 gives:

Corollary 3.8 (Beyond $1/t$ under local dual sharpness for short step and exact line search [LEAN]). *Let $\mathcal{C}$ be compact and convex, and let $f\colon \mathbb{R}^d \to \mathbb{R}$ be smooth and convex. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around the minimizer set $\mathcal{M}$ with some constants (A, q). Then both the global short-step rule and exact line search satisfy*

$$f(x_t) - f^\star = o(1/t).$$

Proof. If for some $x \in \mathcal{C}$ we have $\nabla f(x) = 0$, then x is a global minimizer because of convexity, so that $x \in \mathcal{M}$ and Theorem 3.6 applies. Otherwise ∇f has no zero on the compact set $\mathcal{C}$. Continuity of the gradient then yields $\min_{x \in \mathcal{C}} \|\nabla f(x)\| > 0$, and Proposition 3.7 applies. □

3.2 Open-Loop Step-Sizes

We will now consider the open-loop step-size case. For the sake of exposition we prove the classical $\frac{2}{t+2}$ variant in the main text. For fixed $\ell \ge 2$, the same qualitative argument extends to $\frac{\ell}{t+\ell}$; see Remark A.7 in the appendix. In contrast to the previous arguments for short steps and line search in Section 3.1, we do not have one-step contractions and hence the argument is significantly more involved; we will also not have to distinguish cases according to whether $\nabla f(\mathcal{C}) > 0$.

Theorem 3.9 (LDS + classical open-loop / conditional form [LEAN]). *Let $M \subseteq \mathcal{C}$ be nonempty. Let (x_t) be the Frank–Wolfe iterates generated by the classical open-loop rule $\gamma_t = 2/(t+2)$. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around M with some constants (A, q), where $q \ge 2$. Let $x^\star \in \mathcal{M}$, and let $(p_t) \subseteq M$ satisfy $f(p_t) = f(x^\star)$ for all t. If $\delta_t \stackrel{\text{def}}{=} \|x_t - p_t\| \to 0$, then*

$$(t+2)\big(f(x_t) - f(x^\star)\big) \to 0.$$

Proof deferred to Section A.7 (Deferred Proofs).

Following the proof of Theorem 3.6 we immediately obtain:

Theorem 3.10 (LDS + open-loop [LEAN]). *Assume that $\mathcal{C}$ is compact and convex. Let (x_t) be the Frank–Wolfe iterates generated by the classical open-loop rule $\gamma_t = 2/(t+2)$. Suppose that $\mathcal{C}$*

together with its LMO satisfies local dual sharpness around the minimizer set $\mathcal{M}$ with some constants (A, q), where $q \geq 2$. Then

$$f(x_t) - f^\star = o(1/t).$$

Proof. By Proposition 2.2, $F_t \to 0$. Hence Lemma 2.3 gives

$$\operatorname{dist}(x_t, \mathcal{M}) \to 0.$$

Choose $p_t \in \mathcal{M}$ with

$$\|x_t - p_t\| = \operatorname{dist}(x_t, \mathcal{M}).$$

Then $\|x_t - p_t\| \to 0$. Now apply Theorem 3.9. □

4 Quantitative Rates under a Local Hölder Error Bound

The LDS theorems in Sections 3.1 and 3.2 use only that the iterates approach the minimizer set, which is essentially optimal without further structure on f. With a local Hölder error bound, the same mechanisms become quantitative and yield explicit tail rates after an unoptimized burn-in t_0, generalizing the uniformly-convex-set Hölder-error-bound regime; see Table 1.

Definition 4.1 (Local Hölder error bound of order r)**.** Let $r \geq 2$. We say that f satisfies a *local Hölder error bound of order r* around the minimizer set $\mathcal{M}$ if there exist constants $B, \rho > 0$ such that

$$\operatorname{dist}(x, \mathcal{M})^r \leq B\big(f(x) - f^\star\big) \qquad \forall x \in \mathcal{C} \text{ with } \operatorname{dist}(x, \mathcal{M}) < \rho.$$

Sometimes the above is also written as $\operatorname{dist}(x, \mathcal{M}) \leq c\big(f(x) - f^\star\big)^\theta$, which we recover with parameters $r = 1/\theta$ and $B = c^r$. Thus quadratic growth, and in particular μ-strong convexity, gives $r = 2$ (one may take $B = 2/\mu$), while local order-p growth gives $r = p$. The restriction $r \geq 2$ corresponds to $0 < \theta \leq 1/2$.

Theorem 4.2 (Quantitative rate under LDS + local HEB [LEAN])**.** *Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around the minimizer set $\mathcal{M}$ with constants (A, q), where $q \geq 2$. Assume also that f satisfies a local Hölder error bound of order $r \geq 2$ around $\mathcal{M}$ with constants (B, ρ). Then for the global short-step rule, exact line search, and the classical open-loop rule, there exist $t_0 \in \mathbb{N}$ and $K > 0$ such that $f(x_t) - f^\star \leq \frac{K}{(t - t_0 + 1)^{rq/(rq-2)}}$ for all $t \geq t_0$.*

Proof deferred to Section A.7 (Deferred Proofs).

5 A Computational Example

We consider the feasible region from Example A.3, which satisfies LDS but is not uniformly convex:

$$\mathcal{C}_{\text{stad}} \stackrel{\text{def}}{=} \big([-1, 1] \times \{0\}\big) + B_2(0, 1), \qquad p \stackrel{\text{def}}{=} (2, 0),$$

and the objective family, convex and 1-smooth on $\mathcal{C}_{\text{stad}}$,

$$f_c(x_1, x_2) \stackrel{\text{def}}{=} \frac{1}{2}x_2^2 + \psi(c - x_1), \qquad \psi(u) \stackrel{\text{def}}{=} u - \arctan(u), \qquad c \geq 2.$$

On $\mathcal{C}_{\text{stad}}$ one has $c - x_1 \geq 0$, and the same restriction can be obtained from a global convex smooth extension if desired. For $c = 2$, p is the minimizer and $\nabla f_2(p) = 0$, which is the interesting case. In Figure 1 we depict the resulting primal-gap trajectories for short steps, exact line search, and for the classical open-loop rule from the starting point $x_0 = e_2$, using $L = 1$ in the short-step rule, together with the stadium geometry and the level sets of f_2. The minimizer lies on the rounded cap where we show that LDS holds in the appendix; the flat top and bottom segments show why the stadium as a whole is not uniformly convex and why LDS does not hold for an arbitrary minimizer location. Analogous arbitrarily high-dimensional examples are obtained from capsules or truncated Euclidean balls; we use $\mathbb{R}^2$ only for visualization. A second example with known exponents (a quadratic objective on an ℓ_4 ball with LDS power $q = 4$ and HEB order $r = 2$) is reported in Section A.5; the predicted rate is $O(t^{-4/3})$.

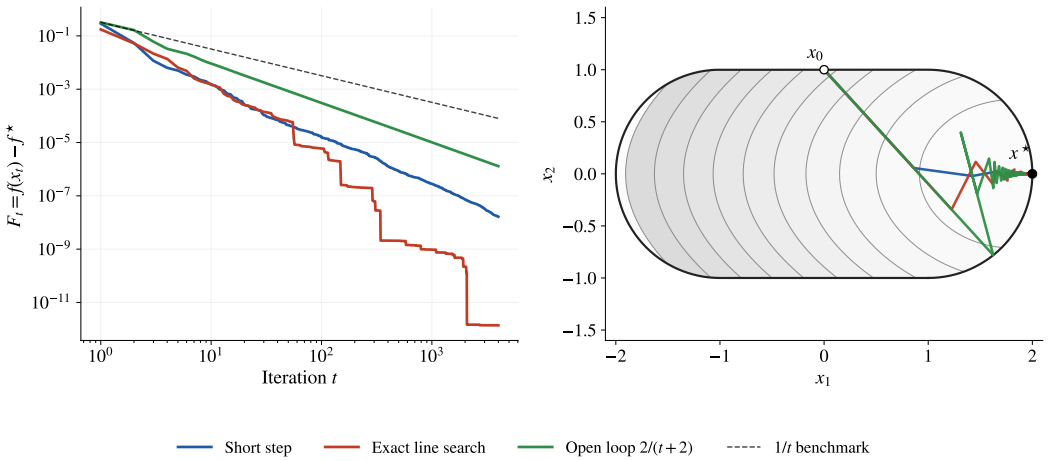


Figure 1: Primal-gap decay and geometry for the stadium. The left panel shows primal gaps for iterations $t \geq 1$ on a log-log plot. The right panel shows the feasible region, the level sets of f_2 restricted to $\mathcal{C}_{\mathrm{stad}}$, and the corresponding Frank–Wolfe trajectories. The finite-horizon curves fall below the displayed $1/t$ guide; the plot is illustrative and does not by itself establish the asymptotic little-o statement.

Limitations. LDS is a condition on a reference region and a fixed LMO selector. In our theorems that region is the objective-dependent minimizer set, so *verifying* LDS may require knowing where the minimizers lie. However, the Frank–Wolfe algorithm automatically adapts when LDS holds, without requiring such verification. Our results are of the form $tF_t \to 0$ for a given LDS instance, not a uniform rate for $\sup_I F_t(I)$ over a class; the associated burn-in is instance-dependent and can be arbitrarily long. Our explicit polynomial rates require an additional objective condition such as a local Hölder error bound and remain tail estimates after an unoptimized burn-in. Nevertheless, a finite-order HEB is not necessary for a quantified nonpolynomial improvement over $1/t$: Proposition A.5 provides an LDS instance without any finite-order HEB whose short-step tail is $F_t \sim L/[t(\log t)^4]$. Flat exposed faces obstruct LDS in standard boundary and low-rank regimes for spectrahedra and nuclear-norm balls.

Acknowledgements and Disclosure of Funding

This research was conducted in the context of the Agentic AI for Mathematics (EF-LiOpt-3) project at the Berlin Mathematics Research Center MATH$^+$ (EXC-2046/2, project ID 390685689), funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy. Significant parts of the auxiliary Lean 4 formalization were produced using an auto-formalization extension of our Agentic Researcher framework [Zimmer et al., 2026]. I thank Christoph Spiegel for independently checking the Lean verification code and for improving both the structure of the Lean code and the guidelines for the auto-formalization. All remaining issues are solely mine.

Competing Interests

The author declares no competing interests.

References

Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditional gradients. In *International conference on machine learning*, pages 735–743. PMLR, 2019.

Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. *Conditional Gradient Methods: From Core Principles to AI Applications*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2025. ISBN 978-1-61197-856-8. doi: 10.1137/1.9781611978568.

M. D. Canon and C. D. Cullum. A tight upper bound on the rate of convergence of Frank–Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968. doi: 10.1137/0306032.

Vladimir F. Demyanov and Alexander M. Rubinov. *Approximate Methods in Optimization Problems*, volume 32 of *Modern Analytic and Computational Methods in Science and Mathematics*. American Elsevier Publishing Company, New York, 1970.

J. C. Dunn. Rates of Convergence for Conditional Gradient Algorithms near Singular and Nonsingular Extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956.

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

Benjamin Grimmer and Ning Liu. Lower bounds for linear minimization oracle methods optimizing over strongly convex sets. *arXiv preprint arXiv:2602.22608*, 2026.

Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986. doi: 10.1007/BF01589445.

Jannis Halbey, Daniel Deza, Max Zimmer, Christophe Roux, Bartolomeo Stellato, and Sebastian Pokutta. Lower Bounds for Frank–Wolfe on Strongly Convex Sets. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 39468–39494. PMLR, 2026. URL https://proceedings.mlr.press/v306/halbey26a.html.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 427–435. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/jaggi13.html.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Projection-Free Optimization on Uniformly Convex Sets. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 19–27, 2021. URL https://proceedings.mlr.press/v130/kerdreux21a.html.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting Frank–Wolfe: Faster rates under Hölderian Error Bounds. *Journal of Optimization Theory and Applications*, 192: 799–829, 2022. doi: 10.1007/s10957-021-01989-7.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Local and Global Uniform Convexity Conditions. In *Data Science and Optimization*, volume 91 of *Fields Institute Communications*, pages 85–120. Springer, 2026. doi: 10.1007/978-3-032-03844-9_4.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. Technical report, Department of Industrial and Systems Engineering, University of Florida, 2013. URL https://optimization-online.org/2013/05/3904/.

Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966. doi: 10.1016/0041-5553(66)90114-5.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *International Conference on Automated Deduction*, pages 625–635. Springer, 2021.

Javier F. Peña. Affine Invariant Convergence Rates of the Conditional Gradient Method. *SIAM Journal on Optimization*, 33(4):2654–2674, 2023.

Sebastian Pokutta. The Frank–Wolfe Algorithm: A Short Introduction. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 126:3–35, 2024. doi: 10.1365/s13291-023-00275-x.

Kazuma K Tsuji, Ken'ichiro Tanaka, and Sebastian Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pages 21864–21883. PMLR, 2022.

Elias Wirth, Thomas Kerdreux, and Sebastian Pokutta. Acceleration of Frank-Wolfe Algorithms with Open-Loop Step-Sizes. In *Proceedings of the 26th International Conference on Artificial*

Intelligence and Statistics (AISTATS), volume 206 of *Proceedings of Machine Learning Research*, pages 77–100. PMLR, 2023. URL https://proceedings.mlr.press/v206/wirth23a.html.

Elias Wirth, Javier Peña, and Sebastian Pokutta. Accelerated Affine-Invariant Convergence Rates of the Frank-Wolfe Algorithm with Open-Loop Step-Sizes. *Mathematical Programming*, 214(1): 201–245, 2025. doi: 10.1007/s10107-024-02180-2.

Philip Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*, pages 1–36. North-Holland, 1970.

Max Zimmer, Nico Pelleriti, Christophe Roux, and Sebastian Pokutta. The Agentic Researcher: A Practical Guide to AI-Assisted Research in Mathematics and Machine Learning. *arXiv preprint arXiv:2603.15914*, 2026.

A Appendix

f	$\mathcal{C}$	minimizer / regime	rule	bound	key
Classical compact-convex regime					
convex	convex	unrestricted	OL / SS / LS	$O(1/t)$	[A]
convex	convex	unrestricted	FW / FO-LMO	$\Omega(1/t)$	[B]
Location-based acceleration					
SC	convex	$x^\star \in \mathrm{Int}(\mathcal{C})$	SS / LS	$O(e^{-rt})$	[C]
convex	UC(q)	LBG, $\inf_{x\in\mathcal{C}} \|\nabla f(x)\| > 0$	SS / LS	$O(e^{-rt})$ for $q=2$; $O(t^{-q/(q-2)})$ for $q>2$	[D]
convex	UC(q)	LBG, $\inf_{x\in\mathcal{C}} \|\nabla f(x)\| > 0$	OL $\ell/(t+\ell)$	$O(t^{-\ell})$ for $q=2$; $O(t^{-\ell+\varepsilon}+t^{-q/(q-2)})$ for $q>2$	[E]
Curved sets without location information					
HEB(θ)	UC(q)	unrestricted	SS / LS	$O(t^{-1/(1-2\theta/q)})$	[F]
HEB(θ)	UC(q)	unrestricted	OL $\ell/(t+\ell)$	$O(t^{-\ell+\varepsilon}+t^{-1/(1-2\theta/q)})$	[G]
SC	SC	unrestricted	SS / LS	$O(1/t^2)$	[H]
SC	SC	unrestricted	SS / LS	$\Omega(1/t^2)$	[I]
SC	SC	unrestricted	deterministic FO-LMO	$\Omega(1/t^2)$	[J]
SC quad.	ℓ_p-ball, $p\geq 3$	zero-gradient boundary minimizer	SS / LS	$\Theta(t^{-p/(p-1)})$	[K]
Wolfe's proper-face regime					
SC	polytope	$x^\star \in$ rel. int(F), $\dim F \geq 1$	SS / LS	$\Omega(1/(t\log^{2+\delta} t))$ i.o. for every $\delta>0$	[L]
HEB(θ)	polytope	$x^\star \in$ rel. int(F), active face identified	OL $\ell/(t+\ell)$	$O(t^{-1/(1-\theta)})$	[M]
Beyond-$1/t$ via local dual sharpness (this paper)					
convex	convex	LDS, unrestricted	SS / LS	pointwise $o(1/t)$	
convex	convex	LDS(q), unrestricted	OL $\ell/(t+\ell)$, $\ell\geq 2$	pointwise $o(1/t)$	
HEB(θ)	convex	LDS(q), unrestricted	SS / LS	$O(t^{-1/(1-2\theta/q)})$	
HEB(θ)	convex	LDS(q), unrestricted	OL $2/(t+2)$	$O(t^{-1/(1-2\theta/q)})$	

Table 2: Known convergence-rate landscape for vanilla Frank–Wolfe. All objective classes in the f-column are assumed convex and smooth, and all feasible sets in the $\mathcal{C}$-column are assumed compact; "SC" means strongly convex, "SC quad." means strongly convex quadratic f, UC(q) means power-type q-uniform convexity, HEB(θ) denotes a (possibly local) Hölderian error bound with exponent $0<\theta\leq 1/2$ (manuscript uses $r=1/\theta$), and LDS denotes local dual sharpness, with LDS(q) indicating the corresponding power q. SS is the global short-step rule, LS is exact line search, OL denotes the displayed open-loop rule or family, and FO-LMO denotes a deterministic first-order method with one linear-minimization-oracle call per iteration. The LDS little-o rows are fixed-instance asymptotic statements. Rows [B], [I], and [J] are uniform worst-case or finite-horizon lower bounds; the remaining rows use the quantifiers stated in their entries and the accompanying discussion. For respective references see keys in Section A.1.

A.1 Related Work and State of the Art (Comprehensive Version)

Table 2 summarizes the vanilla Frank–Wolfe rate landscape most relevant to the present paper. The organizing principle is the mechanism that either improves or obstructs the classical $O(1/t)$ rate. We also separate large-scale oracle lower bounds from smooth low-dimensional witnesses whenever both are available.

Classical compact-convex regime. On a general compact convex set, vanilla Frank–Wolfe with exact line search, global short steps, or the classical open-loop rule has the familiar $O(1/t)$ primal-gap guarantee [A]. The line-search analysis goes back to Frank and Wolfe [1956], while Levitin and Polyak [1966] and Jaggi [2013] give the modern smooth-convex formulation. This baseline is sharp in general: simplex-type examples in Jaggi [2013] already show an $\Omega(1/t)$ obstruction in the low-iteration regime, and Lan [2013] extends the lower-bound picture to deterministic first-order methods that access the feasible region through one LMO call per iteration [B].

Location-based acceleration. A first way to beat $1/t$ is to use information about the position of the minimizer. If the minimizer lies in the interior of the feasible region, then short step and exact line search become linear [C], in a line of work going back to Wolfe [1970, §8] and Guélat and Marcotte [1986]. A second mechanism is the lower-bounded-gradient (LBG) regime on curved sets. When $\inf_{x\in\mathcal{C}} \|\nabla f(x)\| > 0$ and $\mathcal{C}$ is q-uniformly convex, the Frank–Wolfe gap controls the displacement strongly enough to give linear convergence for $q = 2$ and faster-than-$1/t$ polynomial decay for $q > 2$. This picture is classical for strongly convex sets [Demyanov and Rubinov, 1970, Dunn, 1979, Levitin and Polyak, 1966], is presented in modern form by Garber and Hazan [2015], and is extended to uniformly convex sets and affine-invariant growth frameworks by Kerdreux et al. [2021], Peña [2023] [D]. For open-loop rules $\gamma_t = \ell/(t + \ell)$, the same mechanism also yields accelerated rates; the current state of the art is the affine-invariant treatment of Wirth et al. [2025], building on the earlier $\eta_t = 4/(t + 4)$ analysis of Wirth et al. [2023] [E]. Note that rows [C] and [D] are not exhaustive, as they leave out the boundary case in which the unconstrained minimizer lies exactly on $\partial\mathcal{C}$, so that $\nabla f(x^\star) = 0$. In fact, it is this boundary zero-gradient regime that is used by the hard SC/SC instances in [I] and [J].

Curved sets without location information. A different family of results does not assume either an interior minimizer or a gradient bounded away from zero. Here the improvement comes from combining curvature of the feasible region with growth on the objective. In particular, Kerdreux et al. [2021] [F] and Peña [2023] [G] show that a Hölderian error bound together with uniform convexity of the set yields explicit polynomial rates for short step, line search, and open loop. Strongly convex objectives over strongly convex sets fit into this picture and give the familiar $O(1/t^2)$ upper bound of Garber and Hazan [2015] [H]. Recent lower bounds show that this benchmark is essentially sharp in two complementary senses: Halbey et al. [2026] [I] obtain matching smooth small-scale lower bounds on Euclidean balls and ellipsoids, while Grimmer and Liu [2026] [J] prove large-scale lower bounds for deterministic LMO methods. In our notation, the SC/SC regime is a special case of LDS(2) + HEB(1/2) (equivalently $r = 2$), so [I] and [J] match the t^{-2} exponent obtained by specializing Theorem 4.2 in worst-case complexity, although [I] allows the hard initialization and [J] the hard problem instance and dimension to depend on the prescribed iteration horizon. At the low-dimensional end, the explicit quadratic over an ℓ_p-ball of Zimmer et al. [2026] [K] exhibits the slower exponent $\Theta(t^{-p/(p-1)})$ for short step and exact line search at a zero-gradient boundary minimizer.

Wolfe's proper-face regime. When the minimizer lies in the relative interior of a proper face of a polytope, the behavior changes again. Wolfe already observed that short step and exact line search can become much slower in that regime and Canon and Cullum [1968] later refined this observation. A precise formulation of their result is given in Braun et al. [2025, Theorem 2.8]: once an iterate $x_{t_0} \notin F$ with $f(x_{t_0}) < \min_{v\in\mathrm{Vert}(F)} f(v)$ has been reached, for every $\delta > 0$, the primal gap is $\Omega(1/(t \log^{2+\delta} t))$ for infinitely many t [L]. Consequently, no $O(1/(t \log^{2+\eta} t))$ guarantee, with $\eta > 0$, can hold uniformly over this proper-face regime. Open-loop rules behave differently here: once the active face has been identified, the affine-invariant weak-growth framework of Wirth et al. [2025] [M] gives $O(t^{-1/(1-\theta)})$ rates, and for strongly convex objectives this recovers an $O(t^{-2})$ decay.

All results in this related work section concern the vanilla Frank–Wolfe algorithm of [Frank and Wolfe, 1956, Levitin and Polyak, 1966]. Restarted [Kerdreux et al., 2022], away-step/pairwise [Lacoste-Julien and Jaggi, 2015], blended variants [Braun et al., 2019, Tsuji et al., 2022] or otherwise modified variants can exploit additional structural assumptions and yield further acceleration, but those methods are complementary to the present discussion.

A.2 Auxiliary Results

The following inequality is standard; see e.g., [Braun et al., 2025, Kerdreux et al., 2021].

Lemma A.1 (Uniformly-convex geometric gap bound [LEAN])**.** *Let $\mathcal{C}$ be (α, q)-uniformly convex, let $x \in \mathcal{C}$, and let $s \in \operatorname{argmin}_{y\in\mathcal{C}} \langle g, y\rangle$. Then*

$$\frac{\alpha}{2} \|g\| \, \|x - s\|^q \le \langle g, x - s\rangle . \tag{A.1}$$

Proof. By optimality of s,

$$\langle g, y - s \rangle \geq 0 \qquad \forall y \in \mathcal{C}.$$

If $g = 0$, there is nothing to prove. Otherwise, set $z = -g/\|g\|$ and use Definition 2.4 at $\lambda = 1/2$. The point

$$y \stackrel{\text{def}}{=} \frac{x+s}{2} + \frac{\alpha}{4} \|x - s\|^q z$$

belongs to $\mathcal{C}$, hence

$$0 \leq \langle g, y - s \rangle = \frac{1}{2} \langle g, x - s \rangle - \frac{\alpha}{4} \|g\| \, \|x - s\|^q .$$

Rearranging gives the claim. □

A.3 LDS vs. uniform convexity

Proposition A.2 (A uniformly-convex patch with residual gap yields local dual sharpness [LEAN])**.** *Let $M \subseteq \mathcal{C}$ and $K \subseteq \mathbb{R}^d$. Assume that K is (α, q)-uniformly convex for some $q \in \mathbb{R}$ with $q \geq 2$. Suppose there exist $\rho, \beta > 0$ such that whenever $x \in \mathcal{C}$ satisfies*

$$\operatorname{dist}(x, M) < \rho,$$

then $x \in K$, and for every $g \in \mathbb{R}^d$, if s is the atom selected by the LMO for the linear objective $y \mapsto \langle g, y \rangle$, then at least one of the following holds:

1. *$s \in K$ and s also supports K in direction g, namely*

$$\langle g, y - s \rangle \geq 0 \qquad \forall y \in K;$$

2.

$$\beta \|g\| \leq \langle g, x - s \rangle .$$

Then the LMO satisfies local dual sharpness around M with constants

$$\left(\min \left\{ \frac{\alpha}{2}, \frac{\beta}{\max\{1, (\operatorname{diam}(\mathcal{C}))^q\}} \right\}, q \right).$$

Proof. Fix $x \in \mathcal{C}$ with $\operatorname{dist}(x, M) < \rho$, fix $g \in \mathbb{R}^d$, and let s be the atom selected by the LMO. If $g = 0$, then both sides of the desired inequality vanish, so there is nothing to prove.

Assume first that alternative (i) holds. Then $x \in K$ by hypothesis, $s \in K$, and s supports K in direction g. Applying Lemma A.1 to the uniformly convex set K yields

$$\frac{\alpha}{2} \|g\| \, \|x - s\|^q \leq \langle g, x - s \rangle .$$

Assume next that alternative (ii) holds. Because both x and s belong to $\mathcal{C}$, one has

$$D_{\mathcal{C}} \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{C}), \qquad \|x - s\| \leq D_{\mathcal{C}},$$

and since $r \mapsto r^q$ is increasing on the nonnegative reals,

$$\|x - s\|^q \leq D_{\mathcal{C}}^q \leq \max\{1, D_{\mathcal{C}}^q\}.$$

Therefore

$$\frac{\beta}{\max\{1, D_{\mathcal{C}}^q\}} \|g\| \, \|x - s\|^q \leq \beta \|g\| \leq \langle g, x - s \rangle .$$

Taking the minimum of the two coefficients gives

$$\min \left\{ \frac{\alpha}{2}, \frac{\beta}{\max\{1, D_{\mathcal{C}}^q\}} \right\} \|g\| \, \|x - s\|^q \leq \langle g, x - s \rangle ,$$

which is exactly the claimed local dual sharpness bound. □

Example A.3 (Examples beyond globally uniformly convex sets)**.** The local dual sharpness hypothesis is genuinely weaker than global uniform convexity. Two explicit families in $\mathbb{R}^2$ to keep in mind are:

1. the stadium
$$\mathcal{C}_{\mathrm{stad}}(a) \stackrel{\mathrm{def}}{=} \big([-a,a]\times\{0\}\big) + B_2(0,1), \qquad a>0,$$
together with any compact M contained in the relative interior of the right rounded cap, equivalently of the open semicircle
$$\Gamma_a^+ \stackrel{\mathrm{def}}{=} \{(a+\cos\theta,\sin\theta) : -\pi/2<\theta<\pi/2\}, \qquad K_a^+ \stackrel{\mathrm{def}}{=} (a,0)+B_2(0,1);$$

2. the truncated disk
$$\mathcal{C}_{\mathrm{tr}}(b) \stackrel{\mathrm{def}}{=} B_2(0,1)\cap\{x_1\le b\}, \qquad 0<b<1,$$
together with any compact M contained more specifically in the lower arc
$$\Gamma_b^- \stackrel{\mathrm{def}}{=} \{(x_1,x_2)\in\partial B_2(0,1) : x_2<-\sqrt{1-b^2}\}.$$

Neither family is globally uniformly convex: the stadium contains a line segment, while the truncated disk contains a facet. Nevertheless, in both cases the active curved patch has a Euclidean-ball geometry, hence $(\alpha,2)$-uniformly convex, and the complementary LMO directions retain a strictly positive residual support gap. So Proposition A.2 applies and yields local dual sharpness with $q=2$. Higher-dimensional capsules and truncated Euclidean balls satisfy the same patch/residual-gap criterion; the planar examples are used only to visualize the mechanism.

Proof and details for the local-gap examples [LEAN] . We check the hypotheses of Proposition A.2 separately for the two examples.

1. **Stadium.** Let
$$\overline{\Gamma}_a^+ = K_a^+\cap\partial\mathcal{C}_{\mathrm{stad}}(a) = \{(a+\cos\theta,\sin\theta) : -\pi/2\le\theta\le\pi/2\}.$$
Since M is a compact subset of the relative interior of $\overline{\Gamma}_a^+$, there exists $\rho>0$ such that the closed neighborhood
$$U_a \stackrel{\mathrm{def}}{=} \{x\in\mathcal{C}_{\mathrm{stad}}(a) : \mathrm{dist}(x,M)\le\rho\}$$
is contained in K_a^+ and does not meet the two endpoints $(a,\pm1)$. The set K_a^+ is a translate of the Euclidean unit disk, hence $(\alpha,2)$-uniformly convex for some $\alpha>0$.

Fix $x\in U_a$ and $g\in\mathbb{R}^2$. If $g=0$, alternative (ii) is immediate, so assume $g\neq0$ and write $u\stackrel{\mathrm{def}}{=} g/\|g\|$. If the selected atom s lies in K_a^+, then in fact s is the minimizer of the linear functional $y\mapsto\langle u,y\rangle$ over the disk K_a^+, so alternative (i) in Proposition A.2 holds.

It remains to consider the case where the selected atom lies outside K_a^+. This can only happen for directions with $u_1\ge0$. To get a uniform constant, consider the compact closure
$$\Sigma_a \stackrel{\mathrm{def}}{=} \{u\in\mathbb{R}^2 : \|u\|=1,\ u_1\ge0\}.$$
For $u\in\Sigma_a$, the minimum of $y\mapsto\langle u,y\rangle$ over the stadium is
$$m_a(u) = \min_{y\in\mathcal{C}_{\mathrm{stad}}(a)}\langle u,y\rangle = -au_1-1,$$
because $\mathcal{C}_{\mathrm{stad}}(a)=([-a,a]\times\{0\})+B_2(0,1)$. Hence, for every selected atom s,
$$\langle g,x-s\rangle = \|g\|\big(\langle u,x\rangle - m_a(u)\big).$$
Define
$$\Phi_a(x,u) \stackrel{\mathrm{def}}{=} \langle u,x\rangle - m_a(u) \qquad (x\in U_a,\ u\in\Sigma_a).$$
This function is continuous on the compact set $U_a\times\Sigma_a$. If $\Phi_a(x,u)=0$, then x is itself a minimizer of $y\mapsto\langle u,y\rangle$ over $\mathcal{C}_{\mathrm{stad}}(a)$. For $u_1>0$, every minimizer lies on the left rounded cap. For $u_1=0$, every minimizer lies on one of the horizontal line segments together with its two endpoints. Since $U_a\subseteq K_a^+\setminus\{(a,\pm1)\}$, neither possibility can occur. Therefore $\Phi_a>0$ on $U_a\times\Sigma_a$, and compactness gives
$$\beta_a \stackrel{\mathrm{def}}{=} \min_{(x,u)\in U_a\times\Sigma_a}\Phi_a(x,u)>0.$$

Consequently, whenever the selected atom lies outside K_a^+,

$$\langle g, x - s\rangle \geq \beta_a \|g\|,$$

so alternative (ii) in Proposition A.2 holds as well. The proposition therefore yields local dual sharpness with $q = 2$ around M.

2. **Truncated disk.** Write

$$h \stackrel{\text{def}}{=} \sqrt{1 - b^2}, \qquad K \stackrel{\text{def}}{=} B_2(0, 1).$$

Since $M \subseteq \Gamma_b^-$ is compact, there exist $\rho, \eta > 0$ such that

$$U_b \stackrel{\text{def}}{=} \{x \in \mathcal{C}_{\text{tr}}(b) : \operatorname{dist}(x, M) \leq \rho\} \subseteq \{x \in \mathbb{R}^2 : x_2 \leq -h - \eta\}.$$

In particular, U_b is disjoint from the truncating facet

$$F_b \stackrel{\text{def}}{=} \{(b, t) : -h \leq t \leq h\}.$$

The set $K = B_2(0, 1)$ is again $(\alpha, 2)$-uniformly convex for some $\alpha > 0$.

Fix $x \in U_b$ and $g \in \mathbb{R}^2$. As before, the case $g = 0$ is immediate from alternative (ii), so assume $g \neq 0$ and write $u \stackrel{\text{def}}{=} g/\|g\|$. If the full-disk minimizer $-u$ satisfies $-u_1 \leq b$, then $-u \in \mathcal{C}_{\text{tr}}(b)$, so the selected atom is exactly $-u$, which supports the disk K in direction g. Thus alternative (i) holds.

It remains to consider directions for which $-u_1 > b$, equivalently $u_1 < -b$. For a uniform lower bound we again pass to the compact closure

$$\Sigma_b \stackrel{\text{def}}{=} \{u \in \mathbb{R}^2 : \|u\| = 1,\ u_1 \leq -b\}.$$

For every $u \in \Sigma_b$, minimizing $y \mapsto \langle u, y\rangle$ over $\mathcal{C}_{\text{tr}}(b)$ forces the minimizer onto the facet F_b, and the optimal value is

$$m_b(u) = \min_{y \in \mathcal{C}_{\text{tr}}(b)} \langle u, y\rangle = bu_1 - h|u_2|.$$

Hence, for every selected atom s,

$$\langle g, x - s\rangle = \|g\| \big(\langle u, x\rangle - m_b(u)\big).$$

Define

$$\Phi_b(x, u) \stackrel{\text{def}}{=} \langle u, x\rangle - m_b(u) \qquad (x \in U_b,\ u \in \Sigma_b).$$

This function is continuous on the compact set $U_b \times \Sigma_b$. If $\Phi_b(x, u) = 0$, then x is a minimizer of $y \mapsto \langle u, y\rangle$ over $\mathcal{C}_{\text{tr}}(b)$, hence $x \in F_b$. But $U_b \cap F_b = \varnothing$, a contradiction. Therefore $\Phi_b > 0$ on $U_b \times \Sigma_b$, and compactness gives

$$\beta_b \stackrel{\text{def}}{=} \min_{(x,u) \in U_b \times \Sigma_b} \Phi_b(x, u) > 0.$$

Consequently, whenever the selected atom does not support the full disk,

$$\langle g, x - s\rangle \geq \beta_b \|g\|,$$

which is alternative (ii). Applying Proposition A.2 finishes the proof.

□

Remark A.4 (Strict convexity alone is not enough [LEAN])**.** The local dual sharpness hypothesis from Definition 3.1 is a *power-type* statement, so compact strict convexity by itself does not suffice. Indeed, after restricting to a sufficiently small neighborhood and closing the body smoothly elsewhere, one can build a compact strictly convex body in $\mathbb{R}^2$ whose boundary near a support point has the superflat graph

$$y = e^{-1/x^2} \qquad (x \neq 0), \qquad y(0) = 0.$$

For the vertical normal direction, the support gap at a boundary point $(u, e^{-1/u^2})$ is then e^{-1/u^2}, while the chord length is comparable to $|u|$. Hence

$$e^{-1/u^2} = o(|u|^q) \qquad \text{for every fixed } q,$$

so no finite-exponent inequality of the form

$$A \|g\|\ \|x - s\|^q \leq \langle g, x - s\rangle$$

can hold near that support point. Thus compact strict convexity is weaker than the power-type local dual sharpness framework used here.

A.4 An example with $o(1/t)$ convergence without local HEB

A natural question to ask is if $o(1/t)$ convergence under LDS implies some polynomial convergence rate of the form $O(t^{-1-\eta})$ for some $\eta > 0$. We will now show that this is not the case by providing an instance whose primal gap is $\Theta(1/(t(\log t)^4))$ and hence is not $O(t^{-1-\eta})$ for any $\eta > 0$.

Fix $a \in (0, 1/2)$ and define the convex profile

$$\phi(u) \stackrel{\text{def}}{=} \begin{cases} 0, & u \le 0, \\ e^{-1/u}, & 0 < u \le a, \\ e^{-1/a} + e^{-1/a} a^{-2}(u - a), & u > a. \end{cases}$$

We will now show that for the instance

$$\min_{x \in B_2(0,1)} \phi(1 - x_1),$$

which satisfies global LDS but no local HEB of finite order, we obtain a primal-gap convergence of order $\Theta(1/(t(\log t)^4))$.

Proposition A.5 (A logarithmic improvement over $1/t$)**.** *Let $\mathcal{C} = B_2(0,1) \subseteq \mathbb{R}^2$, $p = e_1$, and $f(x) \stackrel{\text{def}}{=} \phi(1 - x_1)$. Let $\widehat{L} > 0$ be any Lipschitz constant for ϕ', and run the global short-step rule with smoothness parameter $L = \widehat{L}$ from $x_0 = (1 - u_0)e_1$, where $u_0 \in (0, a)$. Then $\mathcal{C}$ satisfies global LDS with constants $(1/2, 2)$, f is convex and $\widehat{L}$-smooth with unique constrained minimizer p, and f satisfies no local HEB of any finite order. Nevertheless, its short-step primal gap obeys*

$$\lim_{t \to \infty} \frac{t(\log t)^4}{\widehat{L}} \big(f(x_t) - f(p)\big) = 1.$$

In particular,

$$f(x_t) - f(p) \sim \frac{\widehat{L}}{t(\log t)^4} = o(1/t),$$

but this gap is not $O(t^{-1-\eta})$ for any $\eta > 0$.

Proof. We first verify the basic properties of the instance. On $0 < u < a$,

$$\phi'(u) = \frac{e^{-1/u}}{u^2}, \qquad \phi''(u) = \frac{e^{-1/u}(1 - 2u)}{u^4} \ge 0.$$

The values and first derivatives match at 0 and a; on each of the two complementary intervals, i.e., $(-\infty, 0]$ and $[a, \infty)$, the derivative is constant. Thus ϕ is convex and C^1, and ϕ' is Lipschitz because ϕ'' is bounded on $(0, a)$. As $f(x) = \phi(1 - x_1)$, it follows that f is also convex and smooth. Moreover, every $x \in \mathcal{C}$ has $1 - x_1 \ge 0$, with equality only at $p = e_1$, hence p is the unique constrained minimizer.

Next, we verify LDS for our instance via a direct computation. For $g \ne 0$, write $v = g/\|g\|$; the unit-ball LMO is uniquely $S(g) = -v$. Since $\|x\| \le 1$,

$$\|x - S(g)\|^2 = \|x + v\|^2 \le 2\big(1 + \langle v, x \rangle\big),$$

whereas

$$\langle g, x - S(g) \rangle = \|g\| \big(1 + \langle v, x \rangle\big).$$

Hence $\frac{1}{2}\|g\| \|x - S(g)\|^2 \le \langle g, x - S(g) \rangle$ globally; the case $g = 0$ is immediate.

Now, we verify that $f(x)$ does not satisfy local HEB of any order. To this end, for $x(u) = (1 - u)e_1$ with $u \downarrow 0$, one has $\operatorname{dist}(x(u), \{p\}) = u$ and $f(x(u)) - f(p) = e^{-1/u}$. Therefore, for every finite $r > 0$,

$$\frac{\operatorname{dist}(x(u), \{p\})^r}{f(x(u)) - f(p)} = u^r e^{1/u} \xrightarrow{u \downarrow 0} \infty,$$

which rules out a local HEB of finite order.

It remains to compute the trajectory. If $x_t = (1 - u_t)e_1$ with $0 < u_t < a$ (as holds initially and will be preserved by the update), then we have for the gradient $\nabla f(x_t) = -\phi'(u_t)e_1$, for

the Frank–Wolfe vertex $s_t = S(\nabla f(x_t)) = S(-\phi'(u_t)e_1) = e_1$, for the Frank–Wolfe gap $g_t = \langle \nabla f(x_t), x_t - s_t\rangle = \langle -\phi'(u_t)e_1, (1-u_t)e_1 - e_1\rangle = \langle \phi'(u_t)e_1, u_t e_1\rangle = \phi'(u_t)u_t$, and for the distance $d_t = \|x_t - s_t\| = u_t$. Since $\phi'(0) = 0$ and ϕ' is $\widehat{L}$-Lipschitz, $\phi'(u_t) \le \widehat{L}u_t$ (strictly for $u_t > 0$ in this example). Thus the short step step size is $\gamma_t = \frac{g_t}{Ld_t^2} = \phi'(u_t)/(\widehat{L}u_t)$ and with the update

$$\begin{aligned} x_{t+1} &= x_t - \gamma_t(x_t - s_t) \\ \Leftrightarrow (1-u_{t+1})e_1 &= (1-u_t)e_1 - \gamma_t((1-u_t)e_1 - e_1) \\ \Leftrightarrow u_{t+1} &= u_t - \gamma_t u_t, \end{aligned}$$

and plugging in the definition of γ_t we obtain

$$u_{t+1} = u_t - \frac{e^{-1/u_t}}{\widehat{L}u_t^2}. \tag{A.2}$$

The strict bound $\gamma_t < 1$ implies $0 < u_{t+1} < u_t < a$, so this trajectory form is indeed preserved. This sequence decreases to zero: if it had a positive limit $\bar{u} > 0$, then the decrement in (A.2) would converge to the positive constant $e^{-1/\bar{u}}/(\widehat{L}\bar{u}^2)$, contradicting $u_t \ge 0$.

As before we are going to analyze reciprocals. To this end, set $y_t \stackrel{\text{def}}{=} 1/u_t$ and $b_t \stackrel{\text{def}}{=} \widehat{L}^{-1}e^{-y_t}y_t^3$. Then $b_t \to 0$ as $u_t \to 0$ and the recurrence becomes

$$y_{t+1} = \frac{y_t}{1-b_t} \quad \text{and} \quad y_{t+1} - y_t = \frac{e^{-y_t}y_t^4}{\widehat{L}(1-b_t)},$$

which can be seen as follows: for the first equation write

$$\frac{1}{y_{t+1}} = u_{t+1} = u_t - \frac{e^{-1/u_t}}{\widehat{L}u_t^2} = \frac{1}{y_t} - \frac{e^{-y_t}y_t^2}{\widehat{L}} = \frac{1}{y_t} - \frac{b_t}{y_t} = \frac{1-b_t}{y_t},$$

and for the second equation now observe

$$y_{t+1} - y_t = \frac{y_t}{1-b_t} - y_t = \frac{y_t b_t}{1-b_t} = \frac{e^{-y_t}y_t^4}{\widehat{L}(1-b_t)},$$

via plugging in the definition of b_t in the numerator in the last step. In particular, $y_{t+1} - y_t \to 0$, as $e^{-y_t}y_t^4 \to 0$ for $u_t \to 0$.

Define $R(y) \stackrel{\text{def}}{=} \widehat{L}e^y/y^4$. Since $F_t = e^{-y_t}$ and by the formula for $y_{t+1} - y_t$ above,

$$R(y_t) = \frac{\widehat{L}}{y_t^4 F_t} \quad \text{and} \quad R(y_t) = \frac{1}{(1-b_t)(y_{t+1}-y_t)} \sim \frac{1}{y_{t+1}-y_t}.$$

We now show that the increments of $R(y_t)$ tend to one. Differentiating gives

$$R'(y) = \frac{\widehat{L}e^y}{y^4}\left(1 - \frac{4}{y}\right).$$

For every sufficiently large t, the mean value theorem ensures that there exists some $\xi_t \in (y_t, y_{t+1})$ such that

$$\begin{aligned} R(y_{t+1}) - R(y_t) &= R'(\xi_t)(y_{t+1} - y_t) \\ &= \frac{\widehat{L}e^{\xi_t}}{\xi_t^4}\left(1 - \frac{4}{\xi_t}\right)\frac{e^{-y_t}y_t^4}{\widehat{L}(1-b_t)} \\ &= \frac{e^{\xi_t - y_t}(y_t/\xi_t)^4(1 - 4/\xi_t)}{1-b_t}. \end{aligned}$$

As $\xi_t \in (y_t, y_{t+1})$, we have

$$0 \le \xi_t - y_t \le y_{t+1} - y_t \longrightarrow 0,$$

and hence $e^{\xi_t - y_t} \to 1$. Moreover,

$$0 \le \frac{\xi_t}{y_t} - 1 \le \frac{y_{t+1} - y_t}{y_t} \longrightarrow 0,$$

so $(y_t/\xi_t)^4 \to 1$. Finally, $\xi_t \geq y_t \to \infty$ gives $1 - 4/\xi_t \to 1$, while $b_t \to 0$ gives $(1 - b_t)^{-1} \to 1$. Thus every factor in the formula above tends to one, and hence

$$R(y_{t+1}) - R(y_t) \longrightarrow 1.$$

Telescoping from any fixed sufficiently large T and then dividing by t gives

$$\frac{R(y_t)}{t} = \frac{R(y_T)}{t} + \frac{t-T}{t}\left[\frac{1}{t-T}\sum_{k=T}^{t-1}\big(R(y_{k+1}) - R(y_k)\big)\right] \longrightarrow 1,$$

because the bracketed Cesàro average tends to one and $(t-T)/t \to 1$. Thus

$$\frac{R(y_t)}{t} = \frac{\widehat{L}}{t y_t^4 F_t} \longrightarrow 1,$$

and hence $\frac{t y_t^4 F_t}{\widehat{L}} \to 1$. It remains to show that $y_t/\log t \to 1$, in order to obtain $\frac{t(\log t)^4 F_t}{\widehat{L}} \to 1$. To this end, more explicitly, the preceding positive ratio satisfies $R(y_t)/t = 1 + o(1)$. Applying the logarithm to this ratio we have

$$\log \frac{R(y_t)}{t} = \log(1 + o(1)) = o(1).$$

Using $R(y) = \widehat{L}e^y/y^4$ we can expand the logarithm

$$\log \widehat{L} + y_t - 4\log y_t - \log t = o(1).$$

Since $y_t \to \infty$, division by y_t yields

$$\frac{\log t}{y_t} = 1 - 4\frac{\log y_t}{y_t} + \frac{\log \widehat{L}}{y_t} + o\left(\frac{1}{y_t}\right) \longrightarrow 1,$$

so that $y_t/\log t \to 1$ follows and so

$$\frac{t(\log t)^4 F_t}{\widehat{L}} \longrightarrow 1.$$

Finally, $t^{1+\eta} F_t \sim \widehat{L} t^\eta/(\log t)^4 \to \infty$ for every $\eta > 0$, showing the last assertion. □

Figure 2 displays the two-dimensional level sets and selected Frank–Wolfe iterates alongside the primal tail. Since $f(x) = \phi(1 - x_1)$, the level sets are vertical sections of the ball. The slow convergence of $y_t/\log t$ to one explains the visible finite-time separation from the final $L/(t(\log t)^4)$ equivalent.

Remark A.6 (Dependence on step size rule)**.** For our objective $f(x) = \phi(1-x_1)$, at every nonoptimal point the gradient is a negative multiple of e_1, and therefore the LMO uniquely selects the minimizer p. As there are no other minimizers on the segment $[x_0, p]$, exact line search takes $\gamma_0 = 1$, while the classical open-loop rule has $\gamma_0 = 1$ by definition. Thus both rules reach $x_1 = p$. At the next iteration, the Frank–Wolfe gap satisfies

$$g_1 = \langle \nabla f(p), p - s_1 \rangle = 0,$$

independent of any LMO tie-breaking at the zero gradient. The Frank–Wolfe gap is available in each iteration from the gradient and the already computed LMO atom. Thus, under the standard zero-gap stopping convention, both methods terminate at p before taking a second update; in particular, the Frank–Wolfe algorithm with the open-loop rule does not subsequently move away from the optimum. Hence Proposition A.5 answers the existence question only for global short steps; analogous examples for exact line search and classical open loop remain open but are likely possible to construct.

A.5 An additional example with known q and r

In this section we provide an example for which we can explicitly compute the predicted convergence rate via Theorem 4.2 and compare this to the numerically obtained rates, showing that our estimation is empirically tight on long trajectories.

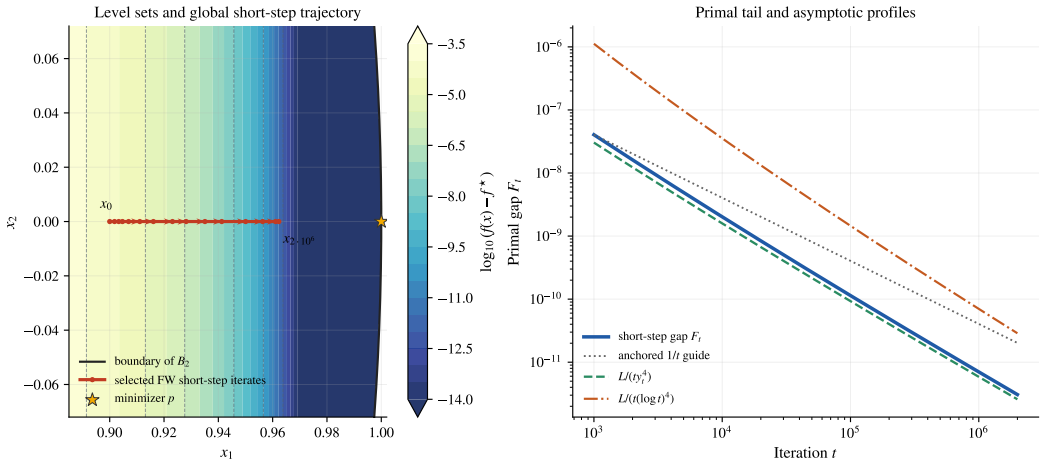


Figure 2: Fixed superflat short-step instance from Proposition A.5, with $a = 1/4$, $u_0 = 0.1$, and $\widehat{L} \approx 2.55$. Left: level sets on a zoom of the right cap of $B_2(0,1)$ and selected iterates of global-short-step Frank–Wolfe; the star marks the minimizer $p = e_1$. Right: the primal gap, an anchored $1/t$ guide, the intermediate profile $\widehat{L}/(ty_t^4)$, and the final equivalent $\widehat{L}/(t(\log t)^4)$. The displayed tail runs from $t = 10^3$ to $t = 2 \cdot 10^6$.

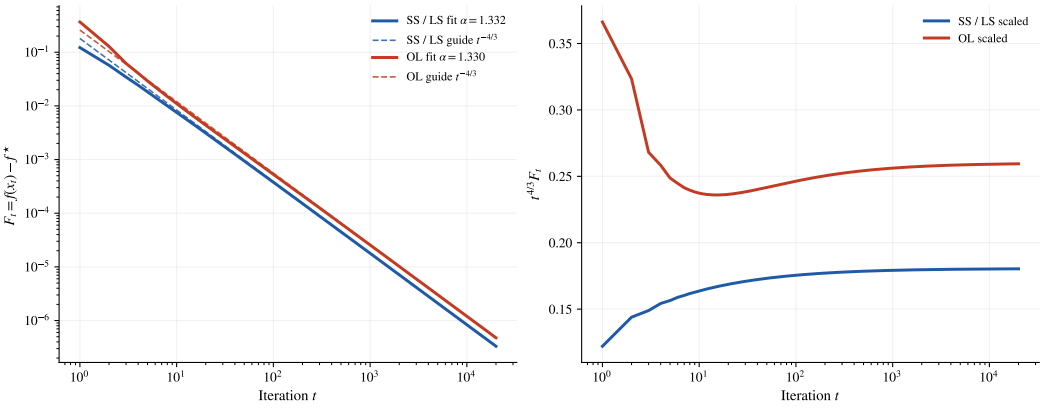


Figure 3: Quadratic ℓ_4-ball example with known LDS power $q = 4$ and HEB order $r = 2$. Left: primal gaps and $t^{-4/3}$ guides. Right: the scaled quantities $t^{4/3}F_t$, which are approximately flat on the tested tail. The plot is a deterministic finite-horizon check of the predicted exponent.

Consider

$$\mathcal{C}_4 \overset{\text{def}}{=} \{x \in \mathbb{R}^2 : \|x\|_4 \le 1\}, \qquad f_4(x) \overset{\text{def}}{=} \frac{1}{2}\|x - e_1\|_2^2 .$$

The set is uniformly convex of power type $q = 4$, hence satisfies LDS(4), and strong convexity of f_4 gives a local HEB of order $r = 2$. Therefore Theorem 4.2 predicts $F_t = O(t^{-rq/(rq-2)}) = O(t^{-4/3})$ for all three step-size rules. Numerically, the fitted exponents are 1.331534 for the short-step trajectory (and the exact-line-search trajectory, since they coincide) and 1.330033 for classical open loop, with $R^2 > 0.999999$ in both cases. Figure 3 shows the corresponding scaled tails. We stress that this is only numerical evidence on a long tail and not an asymptotic verification in the strict sense.

A.6 Generalized Open-Loop Step-Sizes

Remark A.7 (What changes for the fixed $\ell/(t+\ell)$ family)**.** For every fixed integer $\ell \ge 2$ and every LDS power $q \ge 2$, the same minimizer-based open-loop argument under local dual sharpness extends to the rule

$$\gamma_t = \frac{\ell}{t+\ell}.$$

The proof uses the rescaled gaps

$$h_t \stackrel{\text{def}}{=} (t+\ell)F_t$$

instead of $(t+2)F_t$, and the corresponding coefficients become

$$a_t \stackrel{\text{def}}{=} \frac{t(t+\ell+1)}{(t+\ell)^2}, \qquad c_t \stackrel{\text{def}}{=} \frac{t+\ell+1}{(t+\ell)^2}.$$

Accordingly, the curvature and error tails acquire factors of order ℓ^2. The direct shifted-bound branch and the Young-correction branch from the classical open-loop proof persist, with the relevant small-branch and threshold constants enlarged by factors depending only on ℓ. The one-step drop along a dyadic block is therefore weaker by an ℓ-dependent constant factor, but for fixed ℓ it still yields a positive block drop, so the same contradiction argument proves

$$(t+\ell)\big(f(x_t) - f^\star\big) \to 0.$$

The minimizer-set reduction from Theorem 3.10 is unchanged, and Proposition 3.4 immediately yields the uniformly-convex family specialization.

A.7 Deferred Proofs

Lemma 2.5 (Discrete power-law descent [LEAN]). *Let (a_t) be a nonnegative sequence. Assume that for some $0 < \eta \le 1$, $0 < r \le 1$, and some $t_0 \in \mathbb{N}$ with $a_{t_0} \le 1$, it holds $a_{t+1} \le a_t - \eta a_t^{1+r}$ for all $t \ge t_0$. Then $a_t = O\left(t^{-1/r}\right)$.*

Proof. If $a_t = 0$ for some $t \ge t_0$, then the nonnegativity and the recursion force the tail to stay zero, so there is nothing to prove. Otherwise, replacing η by $\bar\eta \stackrel{\text{def}}{=} \min\{\eta, 1/2\}$, we have $\bar\eta a_t^r \le 1/2$ for all $t \ge t_0$. Then

$$a_{t+1} \le a_t(1 - \bar\eta a_t^r), \qquad t \ge t_0.$$

Hence

$$a_{t+1}^{-r} \ge a_t^{-r}(1 - \bar\eta a_t^r)^{-r} \ge a_t^{-r}\big(1 + r\bar\eta a_t^r\big) = a_t^{-r} + r\bar\eta,$$

where we used $\big(1-u\big)^{-r} \ge 1 + ru$ for $0 \le u < 1$. Therefore a_t^{-r} grows at least linearly, so $a_t = O(t^{-1/r})$. □

Proposition 3.7 (LDS + short steps or exact line search when $\nabla f(\mathcal{C}) > 0$ [LEAN]). *Let $\mathcal{M}$ be the minimizer set. Assume that the Frank–Wolfe iterates are generated either by the global short-step rule or by exact line search. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around $\mathcal{M}$ with some constants (A, q) and that $G \stackrel{\text{def}}{=} \min_{x\in\mathcal{C}} \|\nabla f(x)\| > 0$. Then:*

1. *if $q = 2$, then with $\kappa \stackrel{\text{def}}{=} \frac{1}{2}\min\left\{1, \frac{AG}{L}\right\}$, there exists $t_0 \in \mathbb{N}$ such that $F_t \le (1-\kappa)^{t-t_0}F_{t_0}$ for all $t \ge t_0$.*
2. *if $q > 2$, then there exist $t_0 \in \mathbb{N}$ and $K > 0$ such that $F_t \le \frac{K}{(t-t_0+1)^{q/(q-2)}}$ for all $t \ge t_0$.*

Proof. By Proposition 2.2, $F_t \to 0$. Hence Lemma 2.3 gives $\operatorname{dist}(x_t, \mathcal{M}) \to 0$. Let $\rho > 0$ be given by Definition 3.1, and choose t_0 so that $\operatorname{dist}(x_t, \mathcal{M}) < \rho$ for all $t \ge t_0$. Then $A\,\|\nabla f(x_t)\|\; d_t^q \le g_t$ and hence

$$AG\, d_t^q \le g_t \qquad \forall t \ge t_0.$$

Case $q = 2$. Then $d_t^2 \le g_t/(AG)$, so $\frac{g_t^2}{Ld_t^2} \ge \frac{AG}{L}g_t$. Therefore Proposition 2.1 and $F_t \le g_t$ give

$$F_{t+1} \le F_t - \kappa g_t \le (1-\kappa)F_t \qquad \forall t \ge t_0, \qquad \kappa \stackrel{\text{def}}{=} \frac{1}{2}\min\left\{1, \frac{AG}{L}\right\},$$

and iterating yields

$$F_t \le (1-\kappa)^{t-t_0}F_{t_0} \qquad \forall t \ge t_0.$$

Case $q > 2$. Again Proposition 2.1 and $F_t \le g_t$ yield

$$F_{t+1} \le F_t - \frac{1}{2}\min\left\{F_t, \frac{(AG)^{2/q}}{L}F_t^{2-2/q}\right\} \qquad \forall t \ge t_0.$$

By Proposition 2.2, after increasing t_0 if needed we may assume $F_t \le 1$ for all $t \ge t_0$; since $q > 2$, this implies $F_t \ge F_t^{2-2/q}$. Hence

$$F_{t+1} \le F_t - \eta F_t^{2-2/q} \qquad \forall t \ge t_0, \qquad \eta \stackrel{\text{def}}{=} \frac{1}{2}\min\left\{1, \frac{(AG)^{2/q}}{L}\right\},$$

and Lemma 2.5 applied to the shifted tail gives a constant $K > 0$ such that

$$F_t \le \frac{K}{(t - t_0 + 1)^{q/(q-2)}} \qquad \forall t \ge t_0.$$

In either case the tail is asymptotically strictly faster than $1/t$, hence $f(x_t) - f^\star = o(1/t)$. □

Theorem 3.9 (LDS + classical open-loop / conditional form [LEAN]). *Let $M \subseteq \mathcal{C}$ be nonempty. Let (x_t) be the Frank–Wolfe iterates generated by the classical open-loop rule $\gamma_t = 2/(t+2)$. Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around M with some constants (A, q), where $q \ge 2$. Let $x^\star \in \mathcal{M}$, and let $(p_t) \subseteq M$ satisfy $f(p_t) = f(x^\star)$ for all t. If $\delta_t \stackrel{\text{def}}{=} \|x_t - p_t\| \to 0$, then*

$$(t+2)\big(f(x_t) - f(x^\star)\big) \to 0.$$

Proof. For convenience define $h_t \stackrel{\text{def}}{=} (t+2)F_t$, and observe $0 \le h_t \le 2LD^2$ for all $t \ge 1$ by Proposition 2.2; below we always increase burn-in indices so that $t_0 \ge 1$. Let $\rho > 0$ be given by Definition 3.1 and for every t with $h_t > 0$, set $r_t \stackrel{\text{def}}{=} \|p_t - s_t\|$. Since $p_t \in M$, one has $\operatorname{dist}(p_t, M) = 0 < \rho$. Hence convexity together with local dual sharpness at p_t gives $F_t \le \langle \nabla f(x_t), x_t - p_t\rangle$ and $A\,\|\nabla f(x_t)\|\, r_t^q \le \langle \nabla f(x_t), p_t - s_t\rangle$. Adding up the two inequalities we obtain

$$g_t = \langle \nabla f(x_t), x_t - s_t\rangle \ge F_t + A\,\|\nabla f(x_t)\|\, r_t^q.$$

If $F_t > 0$ and hence $\delta_t > 0$, using $F_t \le \|\nabla f(x_t)\|\, \delta_t$ we obtain

$$g_t \ge F_t + A\frac{F_t}{\delta_t} r_t^q.$$

We consider two cases. *Case 1:* If $q = 2$, then $g_t \ge F_t + A\frac{F_t}{\delta_t} r_t^2$. Fix $\varepsilon > 0$ and define

$$a_t \stackrel{\text{def}}{=} \frac{t(t+3)}{(t+2)^2}, \qquad c_t \stackrel{\text{def}}{=} \frac{t+3}{(t+2)^2}, \qquad C_{\mathrm{s}} \stackrel{\text{def}}{=} \frac{4L}{A}.$$

Because $\delta_t \to 0$ and $c_t \to 0$, there exists $t_0 \ge 1$ such that for all $t \ge t_0$,

$$C_{\mathrm{s}}\delta_t < \frac{\varepsilon}{2}, \qquad 4L\delta_t^2 < \frac{\varepsilon}{2}, \qquad 2LD^2 c_t < \frac{\varepsilon}{2}. \tag{A.3}$$

Also $d_t \le \delta_t + r_t$, hence $d_t^2 \le 2\delta_t^2 + 2r_t^2$. Substituting the bounds for g_t and d_t^2 into Proposition 2.1 yields

$$F_{t+1} \le (1 - \gamma_t)F_t + L\gamma_t^2\delta_t^2 + \left(L\gamma_t^2 - A\gamma_t\frac{F_t}{\delta_t}\right) r_t^2. \tag{A.4}$$

The sign of the coefficient of r_t^2 yields two cases. If $L\gamma_t^2 > A\gamma_t F_t/\delta_t$, then

$$F_t < \frac{L}{A}\gamma_t\delta_t = \frac{2L}{A}\frac{\delta_t}{t+2}, \qquad \text{so that} \qquad h_t < \frac{2L}{A}\delta_t \le C_{\mathrm{s}}\delta_t < \frac{\varepsilon}{2}.$$

We call this the small branch. Otherwise the coefficient of r_t^2 is nonpositive; discarding that term from (A.4) and multiplying by $t+3$ gives the refined recurrence

$$h_{t+1} \le a_t h_t + 4L\delta_t^2 c_t. \tag{A.5}$$

Moreover, independently the standard progress estimate gives

$$h_{t+1} \le a_t h_t + 2LD^2 c_t. \tag{A.6}$$

Compared with (A.6), the refined recurrence replaces the fixed diameter error D^2 by the vanishing distance error δ_t^2 and is therefore asymptotically sharper on the tail.

We first prove forward invariance of the sublevel set $\{h_t \le \varepsilon\}$ for $t \ge t_0$. If the small branch holds, then (A.6), $a_t \le 1$, and (A.3) give $h_{t+1} < \varepsilon$. Otherwise (A.5) and $h_t \le \varepsilon$ give

$$h_{t+1} \le a_t \varepsilon + c_t \frac{\varepsilon}{2} \le \varepsilon,$$

because $a_t + c_t/2 \le 1$. This establishes the forward invariance.

Next suppose $h_t > \varepsilon$. The small branch is impossible, and hence (A.5) gives Since $1 - a_t = (t+4)/(t+2)^2$ and $h_t > \varepsilon$, we obtain

$$\begin{aligned} h_t - h_{t+1} &\ge (1 - a_t) h_t - c_t \frac{\varepsilon}{2} \\ &\ge \frac{\varepsilon(t+5)}{2(t+2)^2} \\ &\ge \frac{\varepsilon}{2(t+2)}. \end{aligned}$$

If some arbitrarily late h_n were larger than ε, forward invariance would force $h_t > \varepsilon$ for every $t_0 \le t \le n$. Summing the last inequality from t_0 to $n-1$ would then give

$$h_n \le h_{t_0} - \frac{\varepsilon}{2} \sum_{t=t_0}^{n-1} \frac{1}{t+2},$$

which is negative for all sufficiently large n, contradicting $h_n \ge 0$. Therefore $h_t \le \varepsilon$ eventually. Since $\varepsilon > 0$ was arbitrary, $h_t \to 0$. This proves the theorem when $q = 2$.

Case 2: We therefore assume $q > 2$ below. Fix $\varepsilon > 0$ and define $\beta \stackrel{\text{def}}{=} \frac{2}{q-2}$, $C_\varepsilon \stackrel{\text{def}}{=} \frac{16L}{qA\varepsilon}$, and $R_\varepsilon \stackrel{\text{def}}{=} \frac{q-2}{q} L C_\varepsilon^\beta$. Because $\delta_t \to 0$, both $4L\delta_t^2$ and $4R_\varepsilon \delta_t^\beta$ tend to zero. Hence there exists t_0 such that for all $t \ge t_0$,

$$4L\delta_t^2 + 4R_\varepsilon \delta_t^\beta < \frac{\varepsilon}{2}, \qquad \text{and} \qquad 2LD^2 \frac{t+3}{(t+2)^2} < \frac{\varepsilon}{2}. \tag{A.7}$$

The key point is a dichotomy that we derive in the following. To this end, observe that in the descent estimate the only unknown quantity on the right-hand side is r_t. We isolate its contribution as $\Phi_t(r) \stackrel{\text{def}}{=} L\gamma_t^2 r^2 - A\gamma_t \frac{F_t}{\delta_t} r^q$, with $r \ge 0$. Using $d_t^2 \le 2\delta_t^2 + 2r_t^2$ and $g_t \ge F_t + A(F_t/\delta_t) r_t^q$ in the primal progress estimate from Proposition 2.1 gives, similar to before, $F_{t+1} \le (1-\gamma_t) F_t + L\gamma_t^2 \delta_t^2 + \Phi_t(r_t)$, except we cannot factor out the r_t as before due to mismatching exponents. However, it is enough to obtain a uniform upper bound on $\Phi_t(r)$ for $r \ge 0$. Whenever $F_t > 0$, define $\lambda_t \stackrel{\text{def}}{=} \frac{4L\gamma_t \delta_t}{qAF_t}$, then

$$\Phi_t(r) = L\gamma_t^2 \left(r^2 - \frac{4}{q\lambda_t} r^q \right).$$

Applying Young's inequality to $r^2 = \lambda_t^{-2/q} r^2 \cdot \lambda_t^{2/q}$ with conjugate exponents $q/2$ and $q/(q-2)$ yields $r^2 \le \frac{2}{q\lambda_t} r^q + \frac{q-2}{q} \lambda_t^\beta$ for $r \ge 0$. Hence

$$\Phi_t(r) \le \frac{q-2}{q} L\gamma_t^2 \lambda_t^\beta.$$

This yields the desired dichotomy. Either $\lambda_t > C_\varepsilon \delta_t$, in which case $F_t = \frac{4L\gamma_t \delta_t}{qA\lambda_t} < \frac{4L\gamma_t}{qAC_\varepsilon} = \frac{\varepsilon}{2(t+2)}$, which is equivalent to $h_t < \frac{\varepsilon}{2}$ or else $\lambda_t \le C_\varepsilon \delta_t$, so

$$\Phi_t(r_t) \le \frac{q-2}{q} L\gamma_t^2 C_\varepsilon^\beta \delta_t^\beta = R_\varepsilon \gamma_t^2 \delta_t^\beta.$$

Substituting this into the previous descent estimate gives $F_{t+1} \le (1-\gamma_t) F_t + L\gamma_t^2 \delta_t^2 + R_\varepsilon \gamma_t^2 \delta_t^\beta$ and multiplying by $t+3$ gives

$$h_{t+1} \le a_t h_t + c_t \ \text{coeff}_t, \tag{A.8}$$

where for convenience we introduce $a_t \stackrel{\text{def}}{=} \frac{t(t+3)}{(t+2)^2}$, $c_t \stackrel{\text{def}}{=} \frac{t+3}{(t+2)^2}$, and $\text{coeff}_t \stackrel{\text{def}}{=} 4L\delta_t^2 + 4R_\varepsilon \delta_t^\beta$. Independently, Proposition 2.1 and $d_t \le D$ give the standard bound $h_{t+1} \le a_t h_t + 2LD^2 \frac{t+3}{(t+2)^2}$.

We now prove that the sublevel set $\{h_t \le \varepsilon\}$ is forward invariant for $t \ge t_0$. Assume therefore that $h_t \le \varepsilon$. If the first branch $h_t < \varepsilon/2$ holds, then (A.7) and the standard recurrence give $h_{t+1} < \varepsilon$ because $a_t \le 1$. If instead the recurrence (A.8) holds, then (A.7) implies

$$h_{t+1} \le a_t \varepsilon + c_t \frac{\varepsilon}{2} \le \varepsilon,$$

since $a_t + \frac{c_t}{2} \le 1$.

Next suppose that $t \ge t_0$ and $h_t > \varepsilon$. Then the first branch is impossible, so (A.8) must hold. Using again (A.7), $h_{t+1} \le a_t h_t + c_t \frac{\varepsilon}{2}$ follows. Because $1 - a_t = \frac{t+4}{(t+2)^2}$, we obtain

$$h_t - h_{t+1} \ge \frac{(t+4)h_t}{(t+2)^2} - \frac{\varepsilon}{2}\frac{t+3}{(t+2)^2} \ge \frac{\varepsilon(t+5)}{2(t+2)^2} \ge \frac{\varepsilon}{2(t+2)}.$$

Therefore, on every block $N \le t \le 2N$ with $N \ge t_0$,

$$h_{t+1} \le h_t - \frac{\varepsilon}{4(N+1)}. \tag{A.9}$$

Choose $m \in \mathbb{N}$ so large that $2LD^2 < m\varepsilon/4$, and define $N_k \overset{\text{def}}{=} 2^k(t_0+1) - 1$. If some $n \ge N_m$ satisfied $h_n > \varepsilon$, then the forward invariance would force $h_t > \varepsilon$ for every $t \in \{t_0, t_0+1, \dots, n\}$. Applying (A.9) on each dyadic block $N_k \le t \le 2N_k = N_{k+1} - 1$ gives $h_{N_{k+1}} \le h_{N_k} - \frac{\varepsilon}{4}$ for all $k = 0, \dots, m-1$. By induction, $h_{N_k} \le 2LD^2 - k\frac{\varepsilon}{4}$ for all $k \le m$. At $k = m$ this is negative, contradicting $h_t \ge 0$. Hence $h_t \le \varepsilon$ for all sufficiently large t. Since $\varepsilon > 0$ was arbitrary, $h_t \to 0$ and equivalently $(t+2)F_t \to 0$. □

Theorem 4.2 (Quantitative rate under LDS + local HEB [LEAN]). *Suppose that $\mathcal{C}$ together with its LMO satisfies local dual sharpness around the minimizer set $\mathcal{M}$ with constants (A, q), where $q \ge 2$. Assume also that f satisfies a local Hölder error bound of order $r \ge 2$ around $\mathcal{M}$ with constants (B, ρ). Then for the global short-step rule, exact line search, and the classical open-loop rule, there exist $t_0 \in \mathbb{N}$ and $K > 0$ such that $f(x_t) - f^\star \le \frac{K}{(t-t_0+1)^{rq/(rq-2)}}$ for all $t \ge t_0$.*

Proof. We first consider short steps or exact line search. By Proposition 2.2, $F_t \to 0$, and then Lemma 2.3 gives $\operatorname{dist}(x_t, \mathcal{M}) \to 0$. Choose $p_t \in \mathcal{M}$ with $\delta_t \overset{\text{def}}{=} \|x_t - p_t\| = \operatorname{dist}(x_t, \mathcal{M})$. After increasing t_0 if needed, we may assume that for all $t \ge t_0$, $\delta_t < \rho$, $\delta_t^r \le BF_t$, and $F_t \le 1$. On this tail local dual sharpness and convexity at p_t give

$$A\frac{F_t}{\delta_t} d_t^q \le g_t.$$

For the short-step and exact-line-search part, if $F_t = 0$ at some iteration, the method has reached an optimum and the claimed rate is immediate. We may therefore assume $F_t > 0$ for all t and define $H_t \overset{\text{def}}{=} 1/F_t$ as in the proof of Theorem 3.5. Then Proposition 2.1 gives

$$F_{t+1} \le F_t - \frac{1}{2}\min\left\{g_t, \frac{g_t^2}{Ld_t^2}\right\}.$$

Since $F_{t+1} \le F_t$, this implies

$$H_{t+1} - H_t = \frac{F_t - F_{t+1}}{F_t F_{t+1}} \ge \frac{F_t - F_{t+1}}{F_t^2}.$$

We now split into the same two cases as in Theorem 3.5. If $g_t \ge Ld_t^2$, then $F_t - F_{t+1} \ge \frac{1}{2}g_t \ge \frac{1}{2}F_t$, so

$$H_{t+1} - H_t \ge \frac{1}{2F_t} \ge \frac{1}{2}F_t^{-2/(rq)},$$

because $F_t \le 1$ and $2/(rq) \le 1$.

If instead $g_t \le Ld_t^2$, then the small-gap branch gives $F_t - F_{t+1} \ge \frac{g_t^2}{2Ld_t^2}$, and therefore

$$H_{t+1} - H_t \ge \frac{g_t^2}{2Ld_t^2F_t^2}.$$

Using first $g_t \geq F_t$ and then $A(F_t/\delta_t)d_t^q \leq g_t$, we obtain $H_{t+1} - H_t \geq \frac{1}{2Ld_t^2}$, and $H_{t+1} - H_t \geq \frac{A^2 d_t^{2q-2}}{2L\delta_t^2}$. Taking the weighted geometric mean of these two bounds with weights $(q-1)/q$ and $1/q$, exactly as in Case 2 of Theorem 3.5, yields $H_{t+1} - H_t \geq \frac{A^{2/q}}{2L}\,\delta_t^{-2/q}$. We now combine this with the local Hölder error bound $\delta_t^r \leq BF_t$, to obtain $H_{t+1} - H_t \geq \frac{A^{2/q}}{2L\,B^{2/(rq)}}\,F_t^{-2/(rq)}$. Therefore, for all $t \geq t_0$,

$$H_{t+1} - H_t \geq \eta F_t^{-2/(rq)}, \qquad \eta \stackrel{\text{def}}{=} \min\left\{\frac{1}{2}, \frac{A^{2/q}}{2L\,B^{2/(rq)}}\right\}.$$

Equivalently, $\frac{1}{F_{t+1}} \geq \frac{1+\eta F_t^{1-2/(rq)}}{F_t}$, hence

$$F_{t+1} \leq \frac{F_t}{1+\eta F_t^{1-2/(rq)}}.$$

After increasing t_0 once more, we may assume $\eta F_t^{1-2/(rq)} \leq 1$ on the tail. Using $1/(1+u) \leq 1-u/2$ for $u \in [0,1]$, we obtain

$$F_{t+1} \leq F_t - \frac{\eta}{2}F_t^{2-2/(rq)}.$$

Applying Lemma 2.5 to the shifted tail yields

$$F_t = O\Big(t^{-rq/(rq-2)}\Big),$$

which is the claimed estimate.

For the classical open-loop rule, first assume $q > 2$. Set $h_t \stackrel{\text{def}}{=} (t+2)F_t$, $\sigma \stackrel{\text{def}}{=} \frac{2}{rq-2}$, $\beta \stackrel{\text{def}}{=} \frac{2}{q-2}$, and $\theta \stackrel{\text{def}}{=} \frac{\beta}{r} = \frac{2}{r(q-2)}$. By Proposition 2.2, $0 \leq h_t \leq 2LD^2$ for all $t \geq 1$, and we work after a burn-in $t_0 \geq 1$. The rescaling by $t+2$ factors out the classical $O(1/t)$ rate. Thus the desired estimate $F_t = O(t^{-rq/(rq-2)})$ is equivalent to proving

$$h_t = O(t^{-\sigma}).$$

We prove this by a moving-barrier induction with $\varepsilon_t \stackrel{\text{def}}{=} K(t+2)^{-\sigma}$. The role of K is only to make the barrier high enough after a finite burn-in; it is not optimized.

As above, choose $p_t \in \mathcal{M}$ with $\delta_t \stackrel{\text{def}}{=} \|x_t - p_t\| = \operatorname{dist}(x_t, \mathcal{M})$. Since $F_t \to 0$ and $\operatorname{dist}(x_t, \mathcal{M}) \to 0$, after shifting the index we may assume

$$\delta_t < \rho, \qquad \delta_t^r \leq BF_t = B\frac{h_t}{t+2} \qquad \forall t \geq 0.$$

For every t, repeat the one-step argument from the proof of Theorem 3.9, but with the level ε replaced by $\varepsilon_t \stackrel{\text{def}}{=} \frac{K}{(t+2)^\sigma}$. This yields the same dichotomy: either $h_t \leq \varepsilon_t/2$, or

$$h_{t+1} \leq a_t h_t + c_t\big(4L\delta_t^2 + 4R_t\delta_t^\beta\big),$$

where $a_t \stackrel{\text{def}}{=} \frac{t(t+3)}{(t+2)^2}$, $c_t \stackrel{\text{def}}{=} \frac{t+3}{(t+2)^2}$, $R_t \stackrel{\text{def}}{=} C_0\varepsilon_t^{-\beta}$, and $C_0 > 0$ depends only on (A, L, q). The first alternative is the easy branch: the scaled gap is already well below the barrier. The second alternative is the branch where the LDS correction kicks in; the price is the two error terms involving the distance δ_t to the minimizer set.

We prove by induction that $h_t \leq \varepsilon_t$ for all large t. Assume $h_t \leq \varepsilon_t$. Then the local Hölder error bound gives $\delta_t^r \leq B\frac{\varepsilon_t}{t+2} = \frac{BK}{(t+2)^{1+\sigma}}$, hence

$$\delta_t^2 \leq B^{2/r}K^{2/r}(t+2)^{-2(1+\sigma)/r}, \qquad \delta_t^\beta \leq B^\theta K^\theta (t+2)^{-(1+\sigma)\theta}.$$

Because $r \geq 2$, once $K \geq 1$ we also have $K^{2/r} \leq K$ and $K^{-\beta}K^\theta = K^{-(r-1)\theta} \leq 1$. Consequently there exist constants $C_1, C_2 > 0$, depending only on (A, B, L, q, r), such that on the refined branch

$$h_{t+1} \leq a_t h_t + \frac{C_1 K}{(t+2)^{1+\sigma+\tau}} + \frac{C_2}{(t+2)^{1+\sigma}},$$

where $\tau \stackrel{\text{def}}{=} \frac{2(q-1)}{rq-2} > 0$, as we will argue now: first observe that for the distance error term $c_t\, 4L\delta_t^2$, we have

$$c_t\, 4L\delta_t^2 \le \frac{2}{t+2}\, 4LB^{2/r}K^{2/r}(t+2)^{-2(1+\sigma)/r} \le \frac{C_1 K}{(t+2)^{1+\sigma+\tau}},$$

since $2(1+\sigma)/r - \sigma = \tau$, while for the LDS error term $c_t\, 4R_t\delta_t^\beta$ we have

$$c_t\, 4R_t\delta_t^\beta \le \frac{2}{t+2}\, 4C_0K^{-\beta}(t+2)^{\sigma\beta}B^\theta K^\theta (t+2)^{-(1+\sigma)\theta},$$

and $\theta\big(1-(r-1)\sigma\big) = \sigma$ gives precisely the exponent $1+\sigma$ in the denominator.

Next we will estimate the slack left by the homogeneous part $h_{t+1} \le a_t h_t$ of the scaled open-loop recurrence, in which we will have to fit our two error terms. To this end observe $a_t \le 1 - \frac{1}{t+2}$, while Bernoulli's inequality gives

$$\varepsilon_{t+1} = \frac{K}{(t+3)^\sigma} = \frac{K}{(t+2)^\sigma}\left(1+\frac{1}{t+2}\right)^{-\sigma} \ge \varepsilon_t\left(1-\frac{\sigma}{t+2}\right),$$

hence

$$\varepsilon_{t+1} - a_t\varepsilon_t \ge \frac{(1-\sigma)K}{(t+2)^{1+\sigma}}.$$

Thus, on the refined branch, once K and t_0 are chosen large enough, the two error terms can be absorbed as we will argue now: Choose $K \ge 1$ so large that $C_2 \le \frac{1-\sigma}{4}K$. Then choose t_0 so large that for all $t \ge t_0$, we have

$$\frac{C_1}{(t+2)^\tau} \le \frac{1-\sigma}{4} \qquad \text{and} \qquad 2LD^2\,\frac{t+3}{(t+2)^2} \le \big(1-2^{\sigma-1}\big)\,\varepsilon_{t+1}.$$

Since $\sigma < 1$, the latter is possible because the left-hand side is $O(t^{-1})$ while $\varepsilon_{t+1} = K(t+3)^{-\sigma}$. Enlarging K once more if necessary, we may assume $h_t \le \varepsilon_t$ for all $t \in \{t_0, t_0+1, \dots, 2t_0\}$.

Now fix $t \ge t_0$ and assume $h_t \le \varepsilon_t$. If the small branch holds, then h_t starts with a factor-$1/2$ margin below the barrier, and the standard recurrence gives

$$h_{t+1} \le a_t h_t + 2LD^2\,\frac{t+3}{(t+2)^2} \le \frac{\varepsilon_t}{2} + \big(1-2^{\sigma-1}\big)\,\varepsilon_{t+1} \le \varepsilon_{t+1},$$

because $\varepsilon_t/2 \le 2^{\sigma-1}\varepsilon_{t+1}$. If the refined branch holds, then the barrier slack absorbs the distance and LDS error terms, and the previous two display formulas imply

$$h_{t+1} \le a_t\varepsilon_t + \frac{1-\sigma}{2}\,\frac{K}{(t+2)^{1+\sigma}} \le \varepsilon_{t+1}.$$

Thus $h_t \le \varepsilon_t$ propagates forward for all $t \ge t_0$, and therefore $h_t = O\big(t^{-2/(rq-2)}\big)$. Since $F_t = h_t/(t+2)$, we obtain $F_t = O\big(t^{-rq/(rq-2)}\big)$, as claimed.

It remains to handle the quadratic open-loop branch $q = 2$. As before, set

$$a_t \stackrel{\text{def}}{=} \frac{t(t+3)}{(t+2)^2}, \qquad c_t \stackrel{\text{def}}{=} \frac{t+3}{(t+2)^2}, \qquad C_{\mathrm{s}} \stackrel{\text{def}}{=} \frac{4L}{A}, \qquad B_0 \stackrel{\text{def}}{=} \max\{F_0, LD^2\}.$$

The standard rate gives $0 \le h_t \le 2B_0$ for every t. Repeating the quadratic one-step calculation from Theorem 3.9 gives, on the local-HEB tail, two branches; we refer to the first as the small branch and the second as the refined branch:

$$h_t \le C_{\mathrm{s}}\delta_t \qquad \text{or} \qquad h_{t+1} \le a_t h_t + c_t\, 4L\delta_t^2. \tag{A.10}$$

These two branches hold together with the standard recurrence

$$h_{t+1} \le a_t h_t + 2LD^2 c_t \tag{A.11}$$

and the local error bound $\delta_t^r \le Bh_t/(t+2)$ from our assumption.

Subcase $r > 2$. Let

$$\sigma \stackrel{\text{def}}{=} \frac{1}{r-1} \in (0,1), \qquad S_0 \stackrel{\text{def}}{=} (C_{\mathrm{s}}^r B)^\sigma, \qquad K_0 \stackrel{\text{def}}{=} \max\{1, 2S_0\}.$$

In a first step we choose T beyond the local-HEB burn-in so that for every $t \geq T$,

$$\frac{8LB^{2/r}}{(t+2)^{\sigma}} \leq 1-\sigma, \qquad \frac{8LD^2}{(t+3)^{1-\sigma}} \leq \big(1-2^{\sigma-1}\big)K_0. \tag{A.12}$$

Finally set

$$K \overset{\text{def}}{=} \max\{K_0, 2B_0(T+2)^{\sigma}\}, \qquad \varepsilon_t \overset{\text{def}}{=} \frac{K}{(t+2)^{\sigma}}.$$

Then $h_T \leq \varepsilon_T$ by definition of K: $h_T \leq 2B_0 \leq 2B_0(T+2)^{\sigma}/(T+2)^{\sigma} \leq K/(T+2)^{\sigma} = \varepsilon_T$. This is our induction starting point to prove $h_t \leq \varepsilon_t$ by induction.

In the small branch of (A.10), combining $h_t \leq C_{\mathrm{s}}\delta_t$ with HEB gives for $h_t > 0$

$$h_t^{r-1} \leq \frac{C_{\mathrm{s}}^r B}{t+2}, \qquad \text{hence} \qquad h_t \leq \frac{S_0}{(t+2)^{\sigma}} \leq \frac{\varepsilon_t}{2},$$

where the last inequality uses $K \geq K_0 \geq 2S_0$; for $h_t = 0$ the same bound is immediate. Using this bound together with $a_t \leq 1$ and the standard recursion from (A.11), we get

$$h_{t+1} \leq \frac{\varepsilon_t}{2} + 2LD^2 c_t,$$

and we need to bound the right-hand side by ε_{t+1}. For this we rewrite and bound the curvature term

$$2LD^2 c_t = \frac{2LD^2}{(t+3)^{1-\sigma}} \left(\frac{t+3}{t+2}\right)^2 \frac{1}{(t+3)^{\sigma}} \leq \frac{8LD^2}{(t+3)^{1-\sigma}} \frac{1}{(t+3)^{\sigma}},$$

as $(t+3)/(t+2) \leq 2$. Combining this with the second condition in (A.12) and $K \geq K_0$, we obtain

$$2LD^2 c_t \leq \big(1-2^{\sigma-1}\big)\frac{K}{(t+3)^{\sigma}} = \big(1-2^{\sigma-1}\big)\varepsilon_{t+1}.$$

Moreover,

$$\frac{\varepsilon_t}{2} = \frac{K}{2(t+2)^{\sigma}} = \frac{1}{2}\left(\frac{t+3}{t+2}\right)^{\sigma} \varepsilon_{t+1} \leq 2^{\sigma-1}\varepsilon_{t+1}.$$

Combining the two bounds gives

$$h_{t+1} \leq \frac{\varepsilon_t}{2} + 2LD^2 c_t \leq \varepsilon_{t+1}.$$

In the refined branch, the induction hypothesis, the definition of ε_t, and the HEB imply

$$\delta_t^2 \leq \left(\frac{B\varepsilon_t}{t+2}\right)^{2/r} = \left(\frac{BK}{(t+2)^{1+\sigma}}\right)^{2/r} = \frac{(BK)^{2/r}}{(t+2)^{2(1+\sigma)/r}} = \frac{B^{2/r}K^{2/r}}{(t+2)^{2\sigma}},$$

where the last equality uses $\sigma = 1/(r-1)$, so that $2(1+\sigma)/r = 2\sigma$. Since $K \geq 1$, one has $K^{2/r} \leq K$, and since $c_t \leq 2/(t+2)$, the first condition in (A.12) gives

$$4Lc_t\delta_t^2 \leq \frac{(1-\sigma)K}{(t+2)^{1+\sigma}}.$$

On the other hand,

$$\varepsilon_{t+1} = \varepsilon_t \left(1+\frac{1}{t+2}\right)^{-\sigma} \geq \varepsilon_t \left(1-\frac{\sigma}{t+2}\right)$$

by Bernoulli's inequality. Using $a_t \leq 1 - 1/(t+2)$ therefore yields

$$\varepsilon_{t+1} - a_t\varepsilon_t \geq \frac{(1-\sigma)K}{(t+2)^{1+\sigma}}.$$

Thus the refined recurrence and the induction hypothesis give

$$\begin{aligned} h_{t+1} &\leq a_t h_t + 4Lc_t\delta_t^2 \\ &\leq a_t\varepsilon_t + \frac{(1-\sigma)K}{(t+2)^{1+\sigma}} \\ &\leq a_t\varepsilon_t + \big(\varepsilon_{t+1} - a_t\varepsilon_t\big) = \varepsilon_{t+1}. \end{aligned}$$

Together with the base case and the small-branch estimate, this proves $h_t \le \varepsilon_t$ for every $t \ge T$. Consequently, $h_t = O(t^{-1/(r-1)})$ and $F_t = O(t^{-r/(r-1)})$.

Subcase $r = 2$. At $r = 2$, the exponent $\sigma = 1$ makes the first-order slack used above vanish, so we need to adjust our argument using a shifted barrier. Define

$$C_q \stackrel{\text{def}}{=} 4LB, \qquad s \stackrel{\text{def}}{=} C_q^2 + C_q + 10,$$

and choose T beyond the local-HEB burn-in and large enough that

$$T \ge \lceil 4C_q + 4 \rceil,$$

and further set

$$K_0 \stackrel{\text{def}}{=} \max\{1, 4C_{\mathrm{s}}^2 Bs, 8LD^2(s+1)\}, \qquad K \stackrel{\text{def}}{=} \max\{K_0, 2B_0(T+s)\}, \qquad \varepsilon_t \stackrel{\text{def}}{=} \frac{K}{t+s}.$$

Similar to before, the definition of K initializes the induction:

$$h_T \le 2B_0 \le \frac{K}{T+s} = \varepsilon_T.$$

Assume now that $t \ge T$ and $h_t \le \varepsilon_t$. If we are in the small branch and $h_t > 0$, then the local HEB gives

$$h_t^2 \le C_{\mathrm{s}}^2 \delta_t^2 \le \frac{C_{\mathrm{s}}^2 B}{t+2} h_t,$$

so

$$h_t \le \frac{C_{\mathrm{s}}^2 B}{t+2} \le \frac{C_{\mathrm{s}}^2 Bs}{t+s} \le \frac{K}{4(t+s)} = \frac{\varepsilon_t}{4},$$

where we used $s/(t+s) \ge 1/(t+2)$ because $s \ge 1$ in the second inequality and $K \ge 4C_{\mathrm{s}}^2 Bs$ in the third inequality. The case $h_t = 0$ is again immediate. Since $a_t \le 1$, the standard recurrence yields

$$h_{t+1} \le \frac{\varepsilon_t}{4} + 2LD^2 c_t.$$

The first term fits into half of the next barrier ε_{t+1} because $t + s \ge 1$:

$$\frac{\varepsilon_t}{4} = \frac{K}{4(t+s)} \le \frac{K}{2(t+s+1)} = \frac{\varepsilon_{t+1}}{2}.$$

For the standard curvature term, we use $c_t \le 2/(t+2)$ and $K \ge 8LD^2(s+1)$ to obtain

$$\begin{aligned} 2LD^2 c_t &\le \frac{4LD^2}{t+2} \\ &\le \frac{4LD^2(s+1)}{t+s+1} \\ &\le \frac{K}{2(t+s+1)} = \frac{\varepsilon_{t+1}}{2}. \end{aligned}$$

Therefore the standard recurrence gives $h_{t+1} \le \varepsilon_{t+1}$ in the small branch. If the refined branch holds, then the local HEB gives $\delta_t^2 \le Bh_t/(t+2)$, and consequently

$$\begin{aligned} h_{t+1} &\le a_t h_t + 4Lc_t \delta_t^2 \\ &\le \left(\frac{t(t+3)}{(t+2)^2} + C_q \frac{t+3}{(t+2)^3} \right) h_t. \end{aligned}$$

The choice $s = C_q^2 + C_q + 10$ ensures the elementary coefficient bound

$$\frac{\frac{t(t+3)}{(t+2)^2} + C_q \frac{t+3}{(t+2)^3}}{t+s} \le \frac{1}{t+s+1},$$

whenever $t \ge 4C_q + 4$, which we prove below via a somewhat technical calculation where in essence s as shift is chosen so that the above inequality holds. Since $T \ge 4C_q + 4$, the induction hypothesis therefore gives

$$\begin{aligned} h_{t+1} &\le K \, \frac{\frac{t(t+3)}{(t+2)^2} + C_q \frac{t+3}{(t+2)^3}}{t+s} \\ &\le \frac{K}{t+s+1} = \varepsilon_{t+1}. \end{aligned}$$

The base case and both branches prove $h_t \le K/(t+s) = O(1/t)$ for every $t \ge T$. Hence

$$F_t = \frac{h_t}{t+2} = O(t^{-2}).$$

Combining the two subcases gives

$$F_t = O\Big(t^{-r/(r-1)}\Big),$$

which is exactly $O(t^{-rq/(rq-2)})$ at $q = 2$.

It remains to prove

$$\frac{\frac{t(t+3)}{(t+2)^2} + C_q \frac{t+3}{(t+2)^3}}{t+s} \le \frac{1}{t+s+1},$$

whenever $t \ge 4C_q + 4$. To see this, set

$$\kappa_t \stackrel{\text{def}}{=} \frac{t(t+3)}{(t+2)^2} + C_q \frac{t+3}{(t+2)^3}.$$

Since $t + s > 0$, the desired inequality is equivalent to

$$\kappa_t \le \frac{t+s}{t+s+1} = 1 - \frac{1}{t+s+1}.$$

We first lower-bound the gap to 1. We have

$$\begin{aligned} 1 - \kappa_t &= \frac{t+4}{(t+2)^2} - C_q \frac{t+3}{(t+2)^3} \\ &\ge \frac{1}{t+2} - \frac{2C_q}{(t+2)^2} = \frac{t+2-2C_q}{(t+2)^2}, \end{aligned}$$

where we used $(t+3)/(t+2) \le 2$. It therefore remains to show

$$\frac{t+2-2C_q}{(t+2)^2} \ge \frac{1}{t+s+1}.$$

After cross-multiplication, this is equivalent to

$$\big((t+2) - 2C_q\big)(t+s+1) - (t+2)^2 \ge 0.$$

Using $t + s + 1 = (t+2) + (s-1)$ and keeping $(t+2)$ grouped, the left-hand side is

$$\begin{aligned} &\big((t+2) - 2C_q\big)\big((t+2) + (s-1)\big) - (t+2)^2 \\ &\quad = (t+2)(s - 1 - 2C_q) - 2C_q(s-1) \\ &\quad = (t+2)(C_q^2 - C_q + 9) - 2C_q(C_q^2 + C_q + 9), \end{aligned}$$

where the last equality uses $s = C_q^2 + C_q + 10$. As $t \ge 4C_q + 4$, one has $t + 2 \ge 4C_q + 6$ and moreover $C_q^2 - C_q + 9 > 0$, so that

$$\begin{aligned} &(t+2)(C_q^2 - C_q + 9) - 2C_q(C_q^2 + C_q + 9) \\ &\quad \ge (4C_q + 6)(C_q^2 - C_q + 9) - 2C_q(C_q^2 + C_q + 9) \\ &\quad = 2C_q^3 + 12C_q + 54 > 0. \end{aligned}$$

Consequently,

$$1 - \kappa_t \ge \frac{1}{t+s+1},$$

which gives $\kappa_t \le (t+s)/(t+s+1)$ and therefore proves the claimed coefficient bound. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the LDS condition, the covered step-size rules, the quantitative local-HEB extension, and the scope/limitations of the results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A compact limitations paragraph states the fixed-instance scope, dependence on the minimizer set and fixed LMO selector, the extra local-HEB assumption for explicit rates, the unoptimized burn-in, and the illustrative status of the computations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the theorem statements, proofs are provided in the main text or appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper is theoretical and does not contain main experimental results. Sections A.4, A.5 and 5 contain three deterministic illustrations of the qualitative and quantitative rate mechanisms.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: There are no datasets or main experimental results. The Lean verification will be made available on GitHub, and reusable Lean components will be proposed upstream to mathlib.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper has no training, test split, or randomized experimental protocol; its computations are the deterministic examples in Sections A.4, A.5 and 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: No randomized or statistical experiments are reported. The figure is deterministic.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: There are no experiments. The deterministic illustration is lightweight and CPU-only.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical, uses no human-subject data or sensitive datasets, and does not introduce a deployment or model with safety risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: This is foundational optimization theory work with no direct deployment, dataset, or application-specific societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases no high-risk model, dataset, or scraped content requiring safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The work uses only standard open-source software dependencies and no datasets or pretrained models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the Lean verification, which will be documented and made available on GitHub; reusable Lean components will be proposed upstream to mathlib.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: For transparency, we answer yes: LLMs were used as writing assistants for the manuscript and as coding assistants for the auxiliary Lean development. They were not used as an important, original, or non-standard component of the mathematical method; the authors reviewed all mathematical claims and proofs, and the Lean-marked endpoints were checked by Lean.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.