LOCAL LMO IS SECRETLY A PROJECTION METHOD!

Anonymous authors
Paper under double-blind review

ABSTRACT

The local linear minimization oracle (Ferris & Zavriev, 1996; Richtárik et al., 2026), or Local LMO, solves constrained convex problems without having to compute a projection: it minimizes a linear model over the intersection of the feasible set with a ball around the current iterate. We show that, whenever the ball radius does not exceed the Polyak radius, the Local LMO step is the Euclidean projection of the current iterate onto the intersection of the feasible set with a half-space that separates the iterate from the solution set—a projection that is nonetheless computable by a linear oracle alone. We demonstrate that Local LMO belongs to a broader family of projection methods which may be indexed by the depth of the localizing half-space. For objectives with ϑ-Hölder continuous gradient, every method from this family whose half-space lies sufficiently deep drives the best of its first K iterates to optimality at the universal rate $\mathcal{O}(K^{-(1+\vartheta)/2})$, matching the non-accelerated universal gradient method of Nesterov (2015). Run at the Polyak radius, Local LMO attains the same rate when the constrained optima are also unconstrained ($\|\nabla f(x_\star)\| = 0$), and the rate $\mathcal{O}(K^{-1/(2-\vartheta)})$ otherwise.

1 INTRODUCTION

We study the constrained convex optimization problem

$$\min_{x\in\mathcal{X}} f(x). \tag{1}$$

Problems of the form (1) are ubiquitous in optimization (Boyd & Vandenberghe, 2004) and arise throughout machine learning (Pokutta et al., 2020; Sangalli et al., 2021), statistics and finance (Hathaway, 1985; Hall & Presnell, 1999; Behr et al., 2013) and signal processing (Zhong & Shen, 2026; Mattingley & Boyd, 2010), where constraints encode prior structure, enforce regularization or reflect physical feasibility; in safety-critical deployments, demands for safe, robust and fair models (Donini et al., 2018; Yang et al., 2023) take the same form (Cotter et al., 2019; Dai et al., 2024). There is a rich optimization literature offering a multitude of algorithms for handling such problems. A useful way of mapping out the landscape of these methods is to group them by the oracle used to access the constraint set $\mathcal{X}$. Most approaches fall into one of the following three categories:

1. **Separation oracle.** Such oracles provide, for a given query point x, either a certificate that $x \in \mathcal{X}$ or a hyperplane separating x and $\mathcal{X}$. Classical methods based on this oracle include the ellipsoid method (Grötschel et al., 1993), cutting-plane schemes (Kelley, 1960), and bundle-type epigraph algorithms (Lemaréchal et al., 1995). Membership oracles (Grötschel et al., 1993) are a weaker primitive here, usually converted into separation access.
2. **Projection oracle.** Methods in this class assume access to a projection operator of the form $\mathrm{Proj}^{\phi}_{\mathcal{X}}(x) \in \operatorname{argmin}_{z\in\mathcal{X}} D_\phi(z, x)$, taken with respect to the Bregman divergence of a strictly convex, differentiable function ϕ. This includes *Projected Gradient Descent* (Bertsekas, 1999) when $\phi(x) = \frac{1}{2}\|x\|^2$, mirror descent (Nemirovski & Yudin, 1983) (and its special cases such as exponentiated gradient methods (Kivinen & Warmuth, 1997)), and dual averaging (Nesterov, 2009).
3. **Linear minimization oracle (LMO).** These methods, which include the Frank–Wolfe method (Frank & Wolfe, 1956; Levitin & Polyak, 1966) and its variants, access the constraint set only through linear optimization queries of the form

$$\mathrm{LMO}_{\mathcal{X}}(\nabla f(x)) \in \operatorname*{argmin}_{z\in\mathcal{X}} \langle \nabla f(x), z\rangle\,, \tag{2}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

whenever this minimum is attained. Such oracles are attractive in settings where projections may be expensive or intractable.

The local linear minimization oracle (Ferris & Zavriev, 1996; Richtárik et al., 2026), or Local LMO, falls into the third category and solves constrained convex problems without explicit recourse to a projection oracle. It minimizes a linear model over the intersection of the feasible set $\mathcal{X}$ with a ball around the current iterate. If we denote the current iterate by x_k and the radius of the ball by $t_k > 0$, then the next iterate x_{k+1} computed by Local LMO is

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle , \tag{3}$$

where $\mathcal{B}(x,t) \coloneqq \left\{ z \in \mathbb{R}^d \ : \ \|z - x\| \leq t \right\}$ denotes the Euclidean ball of radius t centered at x. The update (3) *apparently* avoids projections.

To better understand the workings of Local LMO, we shall work under the following standing assumption.

Assumption 1. The set $\mathcal{X} \subseteq \mathbb{R}^d$ is non-empty, closed and convex, the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and the solution set $\mathcal{X}_\star \coloneqq \operatorname{argmin}_{z \in \mathcal{X}} f(z)$ is non-empty.

We write $f_\star \coloneqq \min_{z \in \mathcal{X}} f(z)$ for the optimal function value, and note that the set of optimizers $\mathcal{X}_\star = \mathcal{X} \cap \left\{ z \in \mathbb{R}^d \ : \ f(z) \leq f_\star \right\}$ is then closed and convex.

Our main result, Theorem 2, identifies the Local LMO step (3) with a projection: when the ball radius t_k does not exceed the *Polyak radius*,

$$t(x_k) \coloneqq \frac{f(x_k) - f_\star}{\|\nabla f(x_k)\|}, \tag{4}$$

the Local LMO step (3) is the Euclidean projection of x_k onto the intersection of the feasible set $\mathcal{X}$ with a half-space whose outward normal is the gradient $\nabla f(x_k)$ and which separates the current iterate x_k from the solution set $\mathcal{X}_\star$. A single linear minimization over $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$ thus returns the point a projection oracle would have returned, and the method inherits the properties of a projection method—in particular Fejér monotonicity with respect to $\mathcal{X}_\star$, that is $\|x_{k+1} - x_\star\| \leq \|x_k - x_\star\|$ for every minimizer $x_\star$ (cf. Theorem 2, part 4)—while remaining implementable by a linear oracle. This change of view is what yields the convergence guarantees and universal Hölder rates of Theorem 4. That the half-space contains the solution set $\mathcal{X}_\star$ is what drives our analysis, and it motivates a wider family of projection methods, to which Local LMO belongs and which we present next.

1.1 A FAMILY OF PROJECTION METHODS

We consider methods that, at a point $x_k \in \mathcal{X} \setminus \mathcal{X}_\star$, intersect the feasible set $\mathcal{X}$ with a half-space whose outward normal is the gradient and then project onto the intersection. Since $\nabla f(x) \neq 0$ for all $x \in \mathcal{X} \setminus \mathcal{X}_\star$ (Lemma 5), such a half-space may be described by its depth, that is, the distance from the current iterate to its boundary. For $x \in \mathbb{R}^d$, $g \neq 0$ and $s \geq 0$, we define the half-space

$$\mathcal{H}_x(g, s) \ \coloneqq \ \left\{ z \in \mathbb{R}^d \ : \ \left\langle \tfrac{g}{\|g\|}, x - z \right\rangle \ \geq \ s \right\}, \tag{5}$$

so that $\mathcal{H}_x(g,s)$ has outward normal g and lies at distance s from the point x. For fixed x and g it shrinks as s grows, $\mathcal{H}_x(g, s') \subseteq \mathcal{H}_x(g, s)$ whenever $s \leq s'$, so a larger s means a deeper cut and a smaller set to project onto.

The family of projection methods we consider forms, at the current iterate x_k, the half-space

$$\mathcal{H}_k \ \coloneqq \ \mathcal{H}_{x_k} \left(\nabla f(x_k), \ s_k \right) \tag{6}$$

at an admissible (see (10)) depth s_k, and moves by projecting onto the localized feasible set $\mathcal{X} \cap \mathcal{H}_k$, with $\operatorname{Proj}_{\mathcal{C}}$ the Euclidean projection onto a closed and convex set $\mathcal{C}$,

$$x_{k+1} = \operatorname{Proj}_{\mathcal{X} \cap \mathcal{H}_k}(x_k). \tag{7}$$

We require the cut to exclude the current iterate while retaining every minimizer,

$$x_k \notin \mathcal{H}_k, \qquad \mathcal{X}_\star \subseteq \mathcal{H}_k. \tag{8}$$

The first requirement, $x_k \notin \mathcal{H}_k$, is equivalent to positivity of the chosen depth s_k. The second, $\mathcal{X}_\star \subseteq \mathcal{H}_k$, asks every minimizer to lie at least s_k beyond x_k along the negative gradient direction, and so holds if and only if s_k does not exceed

$$s_\star(x) \;\coloneqq\; \inf_{x_\star \in \mathcal{X}_\star} \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, x - x_\star \right\rangle \tag{9}$$

at $x = x_k$ (Theorem 9), the depth of the deepest localizing cut. The admissible depths at x_k therefore form the interval

$$s_k \;\in\; (\,0,\; s_\star(x_k)\,]\,. \tag{10}$$

Convexity of the objective ensures that this interval is always non-empty. For every $x_\star \in \mathcal{X}_\star$, convexity gives $\langle \nabla f(x), x - x_\star \rangle \geq f(x) - f_\star$, so that

$$\left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, x - x_\star \right\rangle \;\geq\; \frac{f(x) - f_\star}{\|\nabla f(x)\|} \;=\; t(x), \qquad \forall x_\star \in \mathcal{X}_\star, \tag{11}$$

and therefore $s_\star(x) \geq t(x) > 0$. The Polyak radius is itself an admissible depth, and we refer to the cut at that depth as the *Polyak cut*. It is the constrained counterpart of the classical subgradient cut (Polyak, 1969; Censor & Lent, 1982): when $\mathcal{X} = \mathbb{R}^d$, projecting x onto the Polyak cut returns $x - \frac{f(x)-f_\star}{\|\nabla f(x)\|^2}\nabla f(x)$, the gradient step with the Polyak stepsize. Those admissible depths that are at least the Polyak radius, that is $s \in [\,t(x),\; s_\star(x)\,]$, form what we call the *deep-cut band*.

1.2 Contributions

The present work makes the following contributions.

- Local LMO **is a projection method (Theorem 2)**. We show that for every $x \in \mathcal{X} \setminus \mathcal{X}_\star$ and every admissible radius $0 < t \leq t(x)$, where $t(x)$ is the Polyak radius defined in (4), the Local LMO step x_{lmo} of (3) taken at x satisfies

  $$x_{\mathrm{lmo}} \;=\; \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_x(\nabla f(x),\, s_{\mathrm{lmo}}(x,t))}(x), \tag{12}$$

  where the depth $s_{\mathrm{lmo}}(x,t) \coloneqq \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, x - x_{\mathrm{lmo}} \right\rangle$ is admissible. In other words, Local LMO is secretly a projection method!
- **Universal Hölder rate for deep cuts (Theorem 3)**. Every method of the form (7) whose depths s_k lie in the deep-cut band $[\,t(x_k),\; s_\star(x_k)\,]$ drives its best iterate to optimality at rate $K^{-(1+\vartheta)/2}$ on the class of convex functions with ϑ-Hölder continuous gradients, without knowledge of the exponent ϑ or the constant L_ϑ.
- **Universal Hölder rate for** Local LMO **(Theorem 4)**. The Local LMO depth $s_{\mathrm{lmo}}(x_k, t_k)$ never exceeds the Polyak radius and may fall below it, so the previous result does not cover Local LMO in general and we analyze it separately: run at the Polyak radius $t_k = t(x_k)$, it attains the same rate when the constrained minimizers are also unconstrained ones, that is $\|\nabla f(x_\star)\| = 0$ for every $x_\star \in \mathcal{X}_\star$ (Lemma 6), and the rate $K^{-1/(2-\vartheta)}$ otherwise. Both exponents are at least the Frank–Wolfe exponent ϑ (Nesterov, 2018) on the range $[0, 1]$, and neither requires the feasible set $\mathcal{X}$ to be bounded, which the Frank–Wolfe guarantee does.

2 Related Literature

Frank–Wolfe and projection-free methods. The conditional gradient method (Frank & Wolfe, 1956; Levitin & Polyak, 1966) accesses $\mathcal{X}$ only through the global LMO (2) and attains the classical $\mathcal{O}(L\,\mathrm{diam}^2(\mathcal{X})/K)$ rate for L-smooth objectives (Hazan, 2008; Jaggi, 2013; Braun et al., 2025), which cannot be improved by methods restricted to that oracle (Lan, 2013). Under ϑ-Hölder-smoothness the guarantee degrades to $\mathcal{O}(K^{-\vartheta})$ (Nesterov, 2018) and requires $\mathcal{X}$ to be bounded, as the global LMO may otherwise be unbounded below. Away-step, pairwise and fully corrective variants (Wolfe, 1970; Lacoste-Julien & Jaggi, 2015) address the zig-zagging that underlies the $\mathcal{O}(1/K)$ bound but keep the global oracle. For the plain method, linear rates require additional structure, such as a solution interior to $\mathcal{X}$ (Guélat & Marcotte, 1986); other variants attain them over

polytopes instead, at the price of a geometric constant of the feasible set (Lacoste-Julien & Jaggi, 2015). Hazan (2026) avoids projections for smooth convex sets by a supporting-tangent computation rather than a restricted oracle, and Martínez-Rubio & Pokutta (2026) revisit the step size itself; that linear minimization and projection are not interchangeable in cost is made precise by Combettes & Pokutta (2021). Interest in LMO-based methods has been renewed by deep-learning optimizers such as Muon (Jordan et al., 2024), whose update may be read as a linear minimization over a spectral-norm ball, a view developed in Scion (Pethick et al., 2025) and Gluon (Riabinin et al., 2026).

Projection onto a separating cut. The projection step in (7) has a long history. For $\mathcal{X} = \mathbb{R}^d$ it is the subgradient projector of Polyak (1969), taken up for convex feasibility problems by Censor & Lent (1982) and given efficiency estimates by Kiwiel (1996). Projections onto intersections of half-spaces of this kind underlie the weak-to-strong convergence principle of Bauschke & Combettes (2001), and the form $\mathrm{Proj}_{\mathcal{X}\cap\mathcal{H}_k}(x_k)$ appears in the variational-inequality method of Solodov & Svaiter (1999). Closest to the geometry studied here are the level method of Lemaréchal et al. (1995), which with a single cut and level $f_\star$ specializes to the projection onto the Polyak cut, the ballstep subgradient level methods of Kiwiel et al. (1999), which already combine a trust region with a level cut, and the minorant method of Devanathan & Boyd (2024). Fejér monotonicity with respect to $\mathcal{X}_\star$ is the standard engine for this family (Bauschke & Borwein, 1996; Combettes, 2001) and underlies our analysis as well. What seems absent from the literature is the observation that, for an arbitrary closed convex $\mathcal{X}$, a projection of this kind—onto a cut the oracle itself selects—is produced by a single linear minimization over a trust region, which is the content of Theorem 2.

Universal and Hölder-smooth methods. Nesterov (2015) introduced universal gradient methods, attaining $\mathcal{O}(K^{-(1+\vartheta)/2})$ unaccelerated and $\mathcal{O}(K^{-(1+3\vartheta)/2})$ accelerated across the whole Hölder scale without knowledge of (ϑ, L_ϑ), by means of a backtracking line search. The accelerated exponent is optimal and is attained without line search by the bundle-level methods of Lan (2015), at the cost of a compact feasible set and an auxiliary level subproblem per iteration, and by the auto-conditioned method of Li & Lan (2026), which dispenses with compactness (see also Deng et al. (2026)). Among projection-free methods, Ito et al. (2023) remove the need for (ϑ, L_ϑ) at the Frank–Wolfe exponent ϑ through an adaptive line search on a bounded domain; the exponent $\frac{1}{2-\vartheta}$ of Theorem 4 exceeds ϑ on $[0, 1)$ and requires neither a line search nor a bounded domain, at the price of knowing $f_\star$.

Local LMO and the Polyak stepsize. Richtárik et al. (2026) reintroduce the local oracle (3), establish well-posedness, give closed forms for several structured feasible sets, and prove convergence under smoothness, strong convexity and bounded gradients, with the Fejér decrease $\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - t_k^2$ driving the analysis. The oracle is the subproblem of the trust-region successive linear programming scheme analyzed by Ferris & Zavriev (1996), and reappears as the local linear optimization oracle of Garber & Hazan (2016), which relaxes (3) by returning a point within a multiple of the radius and delivers linear convergence over polytopes for smooth, strongly convex objectives. The trust-region viewpoint is shared with the ball-proximal point method of Gruntkowska et al. (2025), which minimizes the objective itself over a ball, and with its non-Euclidean counterpart (Gruntkowska & Richtárik, 2025), in which the linearized ball step studied here appears as the practical surrogate for the exact ball-proximal update. The radius $t(x)$ is the constrained counterpart of the Polyak stepsize (Polyak, 1969), which is known to adapt to problem regularity: Hazan & Kakade (2019) show that a single Polyak-type rule is near-optimal simultaneously across ranges of strong convexity and smoothness, He et al. (2025) give tight analyses together with guarantees under Hölder smoothness and Hölder growth, and Orabona & D'Orazio (2025) recover the rule as gradient descent on a surrogate objective. Like the Polyak stepsize, $t(x)$ presumes knowledge of $f_\star$, which may be estimated online (Abdukhakimov et al., 2025).

3 THEORY

We say that a radius is admissible at x if $0 < t \le t(x)$, where $t(x)$ denotes the Polyak radius defined in (4). Throughout, $g_k \coloneqq \nabla f(x_k)$ and $a_k \coloneqq f(x_k) - f_\star$ denote the gradient and the suboptimality gap at the k-th iterate, respectively, and $R \coloneqq \mathrm{dist}(x_0, \mathcal{X}_\star)$ denotes the initial distance to the solution set.

This section presents three results: Theorem 2 identifies Local LMO as a member of the projection family (7), and thus places every method we consider on one axis, the depth of the cut. Theorem 3 shows that cutting at least as deep as the Polyak cut, that is $s_k \in [\, t(x_k), s_\star(x_k) \,]$, is sufficient to obtain a universal rate. The Local LMO depth never exceeds $t(x_k)$, lying at or below the shallow endpoint of that band, and so Theorem 3 does not cover Local LMO in general; Theorem 4 recovers a universal rate for it through a refined argument. We defer the detailed proofs to the appendix, offering sketches of the key arguments here.

3.1 Local LMO is a projection method

Fix $x \in \mathcal{X} \setminus \mathcal{X}_\star$ and an admissible radius $t \in (0, t(x)]$, and let x_{lmo} solve (3). We define the Local LMO depth

$$s_{\mathrm{lmo}}(x,t) \;:=\; \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, x - x_{\mathrm{lmo}} \right\rangle, \tag{13}$$

that is, the depth of the cut whose boundary passes through the point returned by the oracle. We then obtain the following.

Theorem 2. *Fix $x \in \mathcal{X} \setminus \mathcal{X}_\star$ and an admissible radius $t \in (0, t(x)]$, and let $x_{\mathrm{lmo}} \in \operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle$ be a Local LMO step. Then:*

1. *The Local LMO depth is admissible and does not exceed the radius, that is,*

$$0 \;<\; s_{\mathrm{lmo}}(x,t) \;\le\; t \;\le\; t(x) \;\le\; s_\star(x),$$

   so the half-space $\mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right)$ separates x from, and localizes, the set of minimizers:

$$x \notin \mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right), \qquad \mathcal{X}_\star \subseteq \mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right).$$

2. *The projection onto the localized feasible set (7) coincides with the Local LMO step, that is,*

$$x_{\mathrm{lmo}} = \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_x(\nabla f(x),\, s_{\mathrm{lmo}}(x,t))}(x).$$

   In particular, $\operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle$ is a singleton.

3. *The method takes a step of size t, that is,*

$$\|x_{\mathrm{lmo}} - x\| = \mathrm{dist}\left(x, \mathcal{X} \cap \mathcal{H}_x(\nabla f(x),\, s_{\mathrm{lmo}}(x,t))\right) = t.$$

4. *The squared distance of the new point to any minimizer is reduced by at least t^2, that is,*

$$\|x_{\mathrm{lmo}} - x_\star\|^2 \le \|x - x_\star\|^2 - t^2, \quad \forall x_\star \in \mathcal{X}_\star.$$

Proof sketch. We write $N_{\mathcal{C}}(z)$ for the normal cone of a convex set $\mathcal{C}$ at a point $z \in \mathcal{C}$. The first-order optimality condition for the local subproblem (3) may be restated as $0 \in \nabla f(x) + N_{\mathcal{X}}(x_{\mathrm{lmo}}) + N_{\mathcal{B}(x,t)}(x_{\mathrm{lmo}})$. Since the normal cone of a ball is $\{0\}$ at an interior point and the ray through $x_{\mathrm{lmo}} - x$ at a boundary point (Lemma 8), there exist $n \in N_{\mathcal{X}}(x_{\mathrm{lmo}})$ and $\lambda \ge 0$ with

$$\nabla f(x) + n + \lambda\,(x_{\mathrm{lmo}} - x) \;=\; 0. \tag{14}$$

Parts 3 and 2 both follow from (14), in that order, each by a case split on whether the multiplier λ vanishes; parts 1 and 4 are then consequences. Admissibility is used in the form

$$t \;\le\; t(x) \;=\; \frac{f(x) - f_\star}{\|\nabla f(x)\|} \;\le\; \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, x - x_\star \right\rangle, \qquad \forall x_\star \in \mathcal{X}_\star, \tag{15}$$

which is first-order convexity divided by $\|\nabla f(x)\|$.

Suppose $\lambda > 0$. Since the step is non-zero (Lemma 22), $\lambda(x_{\mathrm{lmo}} - x)$ is a non-zero element of $N_{\mathcal{B}(x,t)}(x_{\mathrm{lmo}})$, and that cone is $\{0\}$ at every interior point of the ball (Lemma 8). Hence, x_{lmo} lies on the sphere and the step has length t. If instead $\lambda = 0$, then $-\nabla f(x) \in N_{\mathcal{X}}(x_{\mathrm{lmo}})$, so $\langle \nabla f(x), x_{\mathrm{lmo}} - x_\star \rangle \le 0$ for every minimizer $x_\star$, while (15), Cauchy–Schwarz and $\|x_{\mathrm{lmo}} - x\| \le t$ give the reverse inequality. The resulting equality forces equality both in the Cauchy–Schwarz bound $\langle \nabla f(x), x_{\mathrm{lmo}} - x \rangle \ge -\|\nabla f(x)\|\,\|x_{\mathrm{lmo}} - x\|$ and in $\|x_{\mathrm{lmo}} - x\| \le t$, so that

$$x - x_{\mathrm{lmo}} \;=\; t\,\frac{\nabla f(x)}{\|\nabla f(x)\|}, \qquad \text{and in particular} \qquad \|x_{\mathrm{lmo}} - x\| \;=\; t, \tag{16}$$

and the step has again length t, establishing part 3.

For part 2, we must establish $\langle x - x_{\mathrm{lmo}}, y - x_{\mathrm{lmo}} \rangle \leq 0$ for every y in the localized feasible set, this being the variational characterization of the projection. First, $\langle n, y - x_{\mathrm{lmo}} \rangle \leq 0$, because $y \in \mathcal{X}$ and $n \in N_{\mathcal{X}}(x_{\mathrm{lmo}})$. Second,

$$y \in \mathcal{H}_x\left(\nabla f(x),\, s_{\mathrm{lmo}}(x,t)\right) \qquad \Longleftrightarrow \qquad \langle \nabla f(x), y - x_{\mathrm{lmo}} \rangle \;\leq\; 0, \tag{17}$$

since inserting the depth (13) into the definition (5) of the half-space reads $\langle \nabla f(x), x - y \rangle \geq \langle \nabla f(x), x - x_{\mathrm{lmo}} \rangle$. For $\lambda > 0$, taking the inner product of (14) with $y - x_{\mathrm{lmo}}$ and discarding the non-positive terms leaves $\lambda \langle x_{\mathrm{lmo}} - x, y - x_{\mathrm{lmo}} \rangle \geq 0$, which is the claim. For $\lambda = 0$, (16) gives $x - x_{\mathrm{lmo}}$ as a positive multiple of $\nabla f(x)$, and the claim is then (17) again.

In part 1, $s_{\mathrm{lmo}}(x,t) \leq t$ is Cauchy–Schwarz applied to the step $x - x_{\mathrm{lmo}}$, whose length is t by part 3, and $s_{\mathrm{lmo}}(x,t) > 0$ holds because the feasible point $(1-\mu)x + \mu x_\star$ with $\mu = t/\|x - x_\star\|$ strictly improves the linear model (Lemma 22). Taking the infimum over $x_\star$ in (15) gives the remaining inequalities $t \leq t(x) \leq s_\star(x)$, which is what makes the cut localize $\mathcal{X}_\star$ (Theorem 9). Part 4 is the projection inequality of part 2 evaluated at a minimizer, which lies in the localized feasible set by part 1, together with the step length (Theorem 10). □

The step length in part 3 and the decrease in part 4 are Theorem 3.1(ii)–(iii) of Richtárik et al. (2026); we rederive them because the projection identity yields them at no extra cost and because Theorem 26 later sharpens part 4. Parts 1 and 2, and the identification of that length with the distance to the localized feasible set, are novel insights. Theorem 2 places Local LMO inside the family (7) and shows that Local LMO is (secretly) a projection method! Two parts carry the analysis that follows: part 4 is Fejér monotonicity (Theorem 10), which makes the squared step lengths summable, and part 1 places Local LMO on the shallow side of the Polyak depth via $s_{\mathrm{lmo}}(x,t) \leq t \leq t(x)$. The next subsection turns to the members that cut at least as deep as the Polyak cut.

3.2 A UNIVERSAL RATE FOR DEEP CUTS

The rates in this subsection and the next are stated on the Hölder scale. We say that the gradient map ∇f of the objective function is Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ (or simply, ϑ-Hölder continuous) if the Bregman divergence $D_f(u,v) \coloneqq f(u) - f(v) - \langle \nabla f(v), u - v \rangle$ satisfies

$$D_f(u,v) \;\leq\; \tfrac{L_\vartheta}{1+\vartheta} \|u - v\|^{1+\vartheta}, \qquad \forall u, v \in \mathbb{R}^d. \tag{18}$$

The exponent $\vartheta = 1$ corresponds to the smooth case, $\vartheta = 0$ to that of bounded gradient variation, and the scale interpolates between them. Following Nesterov (2015), we call a method *universal* if it needs no a priori knowledge of (ϑ, L_ϑ); the methods below are universal in this sense.

For the subfamily of projection methods (7) in the deep-cut band $[t(x), s_\star(x)]$, we obtain the following.

Theorem 3 (Universal Hölder rate on the deep-cut band)**.** *Assume ∇f is Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ as in* (18)*. Let $\{x_k\}$ be generated from $x_0 \in \mathcal{X} \setminus \mathcal{X}_\star$ by $x_{k+1} = \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_{x_k}(\nabla f(x_k),\, s_k)}(x_k)$ with any (possibly adaptive) depths in the deep-cut band, that is, satisfying $t(x_k) \leq s_k \leq s_\star(x_k)$ for all k, the run stopping if some $x_k \in \mathcal{X}_\star$. Then, for every $K \geq 1$, we have*

$$\min_{0 \leq k \leq K} \big(f(x_k) - f_\star\big) \leq \frac{L_\vartheta}{1+\vartheta} \left(\frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}\right)^{\frac{1+\vartheta}{2}}. \tag{19}$$

Proof sketch. The argument rests on three parts, and the Hölder data enter only in the last. First, cutting at least as deep as the Polyak cut: If x_{k+1} lies in the Polyak cut at x_k, then the linearization of f at x_k, evaluated at x_{k+1}, is at most $f_\star$. Subtracting it from $f(x_{k+1})$ leaves

$$f(x_{k+1}) - f_\star \;\leq\; D_f(x_{k+1}, x_k), \tag{20}$$

so the function gap may be upper-bounded by the Bregman divergence of a single step (Lemma 15). Second, Fejér monotonicity (Theorem 10): any cut retaining $\mathcal{X}_\star$ gives $\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \|x_{k+1} - x_k\|^2$, so the squared step lengths sum to at most R^2 and the shortest of K steps obeys $\min_{k<K} \|x_{k+1} - x_k\|^2 \leq R^2/K$. Third, Hölder continuity is invoked once, at that

short step: $D_f(x_{k+1}, x_k) \le \frac{L_\vartheta}{1+\vartheta}\|x_{k+1} - x_k\|^{1+\vartheta}$. Combining the three gives (19). Neither ϑ nor L_ϑ is used to generate the iterates; they appear only in this last line, which is what makes the rate universal. □

The family of projection methods described in Theorem 3 attains the rate $K^{-(1+\vartheta)/2}$ across the whole Hölder scale, the rate of the universal primal gradient method of Nesterov (2015). There, universality is obtained by a backtracking search for a local Hölder constant; here, it comes from the geometry of the cut instead.

3.3 A UNIVERSAL RATE FOR LOCAL LMO

By part 1 of Theorem 2, Local LMO cuts at depth at most $t(x_k)$, that is, at or below the shallow endpoint of the deep-cut band from Theorem 3, and is therefore not covered by the previous result in general.

Theorem 4 (Universal Hölder rate for Local LMO)**.** *Let ∇f be Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ, and let $\{x_k\}$ be generated by* Local LMO *at the Polyak radius $t_k = t(x_k)$ from $x_0 \in \mathcal{X} \setminus \mathcal{X}_\star$, the run stopping if some $x_k \in \mathcal{X}_\star$. Then, for all $K \ge 1$, we have*

$$\min_{0\le k\le K}\big(f(x_k) - f_\star\big) \;\le\; \max\Big\{\, \Lambda_1\, K^{-\frac{1+\vartheta}{2}} \;,\; \Lambda_2\, K^{-\frac{1}{2-\vartheta}} \,\Big\}, \tag{21}$$

where $R = \operatorname{dist}(x_0, \mathcal{X}_\star)$ and $G_\star = \|\nabla f(x_\star)\|$ for any $x_\star \in \mathcal{X}_\star$, and

$$\Lambda_1 \coloneqq \tfrac{25}{2}\, L_\vartheta R^{1+\vartheta}, \qquad \Lambda_2 \coloneqq 25\max\Big\{\big(L_\vartheta R^2 G_\star^{1-\vartheta}\big)^{\frac{1}{2-\vartheta}},\; G_\star R\Big\}. \tag{22}$$

Proof sketch. The first ingredient of the previous argument is unavailable: the Local LMO cut is no deeper than the Polyak cut, so x_{k+1} need not lie in the latter and (20) may fail. The shortfall in depth, weighted by the gradient norm, is what we call the overshoot,

$$\Gamma_k \;\coloneqq\; \|g_k\|\,(t(x_k) - s_{\mathrm{lmo}}(x_k, t_k)) \;\ge\; 0, \tag{23}$$

which vanishes precisely when the Local LMO cut is the Polyak cut (Lemma 23). The exact identity $f(x_{k+1}) - f_\star \;=\; D_f(x_{k+1}, x_k) + \Gamma_k$ shows that the amount by which (20) fails is exactly the overshoot. Since $D_f \ge 0$, no Local LMO step can push the suboptimality below Γ_k, however smooth f may be (Lemma 24).

The quantity that our analysis rests on is, with $\ell_k \coloneqq \langle g_k, x_k - x_{k+1}\rangle$ the decrease of the linear model,

$$\rho_k \;\coloneqq\; \frac{\ell_k}{a_k} \;\in\; (0,1], \tag{24}$$

the fraction of the first-order decrease available at x_k that the oracle actually captures; since $\ell_k = \|g_k\| s_{\mathrm{lmo}}(x_k, t_k)$ and $a_k = \|g_k\| t(x_k)$, the range $(0,1]$ is part 1 of Theorem 2 and $\Gamma_k = (1-\rho_k)a_k$. At the Polyak radius, ρ_k is the cosine of the angle between g_k and the step $x_k - x_{k+1}$: it equals 1 when the step runs straight down the gradient and tends to 0 when the feasible set forces the step almost orthogonal to g_k.

A small ρ_k is bad for the function value—the floor Γ_k leaves $a_{k+1} \ge (1-\rho_k)a_k$, so one step removes at most a ρ_k fraction of the gap, while the descent inequality delivers only $a_{k+1} \le a_k(1 - \rho_k + \delta_k)$, with $\delta_k \coloneqq \frac{L_\vartheta}{1+\vartheta} t_k^{1+\vartheta}/a_k$ the Hölder error—but it is good for the distance. The projection identity of Theorem 2 sharpens Fejér monotonicity to

$$\|x_{k+1} - x_\star\|^2 \;\le\; \|x_k - x_\star\|^2 - t_k^2\,\frac{2-\rho_k}{\rho_k} \tag{25}$$

(Theorem 26), in which the overshoot amplifies the contraction by the factor $(2-\rho_k)/\rho_k \ge 1/\rho_k$: a nearly tangential step barely lowers f, but pulls the iterate closer to $\mathcal{X}_\star$ than its length alone would suggest. Summing gives the refined budget $\sum_{k<K} t_k^2/\rho_k \le R^2$. The two fit together because

$$\underbrace{\frac{\rho_k}{a_k^2}}_{\text{gain in } \Phi_k \coloneqq a_k^{-2}} \cdot \underbrace{\frac{t_k^2}{\rho_k}}_{\text{Fejér cost}} \;=\; \frac{t_k^2}{a_k^2} \;=\; \frac{1}{\|g_k\|^2} \tag{26}$$

does not depend on ρ_k at all: a step that makes little progress in function value pays for it in the Fejér budget by exactly the reciprocal factor, so no choice of ρ_k is free.

The remainder of the proof rests on an accounting argument. We write $\bar{a}$ for the level below which $G_\star$ dominates the Hölder part of the gradient bound; above it the refined budget alone delivers the exponent $\frac{1+\vartheta}{2}$. Below $\bar{a}$, a two-sided gradient bound pins $\|g_k\|$ to within a constant factor of $G_\star$, and the K steps are counted in three groups: those above $\bar{a}$ are few by the budget; those that stall, in the sense $\rho_k < 2\delta_k$, are few because each consumes a fixed share of the budget; and the productive remainder is counted by Cauchy–Schwarz against Φ_k using (26), a count needed only once K is large. This produces the exponent $\frac{1}{2-\vartheta}$, and (21) is the larger of the two bounds. □

Two remarks on (21) are in order. First, the second term disappears when the gradient vanishes at the solution: if $G_\star = 0$, so that every constrained optimum is also an unconstrained one, then (21) holds with Λ_2 replaced by 0, and Local LMO attains the exponent $\frac{1+\vartheta}{2}$ of Theorem 3, differing only in the constant, although its cut is never deeper than the shallow endpoint of the band.

Second, at $\vartheta = 0$ the two exponents in (21) coincide, at $\frac{1}{2}$, and a direct argument improves the constant: a Hölder-0 bound forces $\|g_k\| \le G_\star + L_0$ along the whole trajectory, and Fejér monotonicity alone then gives

$$\min_{0\le k\le K}\big(f(x_k) - f_\star\big) \;\le\; \frac{(G_\star + L_0)\,R}{\sqrt{K}}, \tag{27}$$

the same rate as (21) with a smaller constant (Proposition 30).

4 Experiments

To illustrate the qualitative difference between Local LMO and the deep-cut methods, we consider the two-dimensional objective function

$$f(x, y) = \frac{L}{1+\vartheta}|x|^{1+\vartheta} + Gy \tag{28}$$

to be optimized over the half-space $\mathcal{X} = \{(x, y) \,:\, y \ge 0\}$, where $G \ge 0$ and $L > 0$ are constants and $\vartheta \in (0, 1]$ governs Hölder continuity as defined in (18); the endpoint $\vartheta = 0$ is excluded to ensure differentiability of f.

For $G > 0$, the minimizer is unique, $x_\star = (0, 0)$, and lies on the boundary of the feasible set $\mathcal{X}$; for $G = 0$ the solution set is the ray $\{(0, y) \,:\, y \ge 0\}$, of which $(0, 0)$ is the point nearest to the initial point $x_0 = (R, 0)$, for some $R > 0$, located on the boundary of $\mathcal{X}$. For $G > 0$ the affine part of the objective pushes against the constraint $y \ge 0$, and for every $G \ge 0$ we have $\nabla f(x_\star) = (0, G)$, so $G_\star = \|\nabla f(x_\star)\| = G$. The ϑ-Hölder continuous part has its Hölder singularity at $x = 0$, and (18) holds with $L_\vartheta = 2^{1-\vartheta}L$, independent of G; varying G therefore varies how constrained the optimizer is while ϑ and L_ϑ stay fixed. The two coordinates thus separate the quantities governing Theorems 3 and 4: the Hölder term governs the growth of f along the face $y = 0$, on which the iterates travel, while the affine term contributes nothing to the value there.

We run Local LMO at the Polyak radius $t_k = t(x_k)$ and, for comparison, the projection method (7) at the Polyak cut $s_k = t(x_k)$, the shallow endpoint of the deep-cut band. Both iterations remain on the face $y = 0$, so, writing $x_k = (u_k, 0)$ with $u_k > 0$ and using $a_k = \frac{L}{1+\vartheta}u_k^{1+\vartheta}$ and $g_k = \big(Lu_k^\vartheta,\, G\big)$, the iterates are available in closed form:

$$u_{k+1}^{\mathrm{lmo}} \;=\; u_k - \frac{a_k}{\|g_k\|}, \qquad \|g_k\| = \sqrt{L^2 u_k^{2\vartheta} + G^2}, \qquad \text{whereas} \qquad u_{k+1}^{\mathrm{proj}} \;=\; \frac{\vartheta}{1+\vartheta}\,u_k. \tag{29}$$

The Local LMO update is the step of length t_k supplied by part 3 of Theorem 2, the oracle returning the point $(u_k - t_k, 0)$ on the face for every $G \ge 0$; the projection update, on the other hand, does not involve G at all.

For large G, the Local LMO trajectory in Figure 1 is essentially flat over a long initial phase whose length grows in proportion to G, while the projection method converges linearly and is unaffected. The mechanism is best understood through (29): at a given position u_k the suboptimality a_k does not depend on G, because the affine term vanishes on the face, whereas the gradient norm does. The

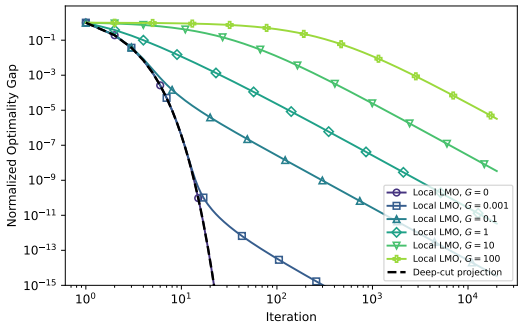

Figure 1: Normalized optimality gap a_k/a_0 on (28) with $\vartheta = 1/2$ and $L = R = 1$: Local LMO at the Polyak radius for a range of values of G, and the projection method (7) at the Polyak cut, which does not depend on G. Local LMO coincides with it at $G = 0$; for larger G, it first traverses a plateau whose length grows linearly in G.

Polyak radius $t_k = a_k/\|g_k\| \le a_k/G$ is therefore throttled, and since the step length equals t_k and a_k decreases here, the iterate advances by at most a_0/G per iteration. Covering the initial distance R therefore takes at least GR/a_0 iterations, which is the plateau one observes in Figure 1.

In (28), the Local LMO step moves along the face of the feasible set, so the fraction (24) is

$$\rho_k \;=\; \frac{L u_k^{\vartheta}}{\sqrt{L^2 u_k^{2\vartheta} + G^2}} \;=\; \frac{\langle g_k, x_k - x_{k+1}\rangle}{\|g_k\|\,\|x_k - x_{k+1}\|}, \tag{30}$$

and a large G tilts the gradient towards the normal of the constraint, driving ρ_k to zero. The projection method is not throttled in this way: it moves the full $\operatorname{dist}(x_k, \mathcal{X} \cap \mathcal{H}_k)$ to the Polyak cut, which here equals t_k/ρ_k, whereas Local LMO moves only $t_k = \operatorname{dist}(x_k, \mathcal{H}_k)$. This slack separates Theorem 3 from Theorem 4. It vanishes when $G = 0$, where the lowest curve of Figure 1 shows the two methods coinciding: the gradient is parallel to the face, the overshoot is zero (Lemma 23), and both updates reduce to $u_{k+1} = \frac{\vartheta}{1+\vartheta} u_k$.

5 Conclusion

We have shown that Local LMO—a single linear minimization over the feasible set intersected with a ball around the current iterate introduced as a projection-free approach—does, in fact, implement a projection. As long as the ball radius does not exceed the Polyak radius $(f(x_k) - f_\star)/\|\nabla f(x_k)\|$, the point the oracle returns is the Euclidean projection of x_k onto the feasible set intersected with a half-space whose outward normal is the gradient and which separates x_k from every minimizer. Local LMO is therefore a projection method that happens to be implementable by a linear oracle, and the half-space it projects onto is one it selects itself. Local LMO is part of the family of projection methods (7) which may be indexed by the depth of the half-spaces.

Cutting at least as deep as the Polyak cut, while still retaining every minimizer, is enough for the rate $K^{-(1+\vartheta)/2}$ on convex objectives with ϑ-Hölder continuous gradient, and the guarantee holds for the whole family of such cuts. Local LMO may cut shallower than the Polyak cut, and the shortfall in depth separates it from the deep-cut methods. Charging the shortfall to the distance to the solution set rather than to the function value sharpens the Fejér contraction instead of destroying it, and yields a universal rate for Local LMO, as well: the same $K^{-(1+\vartheta)/2}$ when the gradient vanishes at the solution, and $K^{-1/(2-\vartheta)}$ otherwise. Both exponents are at least the Frank–Wolfe exponent ϑ over the whole range $[0, 1]$, and neither guarantee requires the feasible set to be bounded.

REFERENCES

Farshed Abdukhakimov, Cuong Anh Pham, Samuel Horváth, Martin Takáč, and Slavomır Hanzely. Polyak stepsize: Estimating optimal functional values without parameters or prior knowledge. *arXiv preprint arXiv:2508.17288*, 2025.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

Heinz H. Bauschke and Patrick L. Combettes. A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces. *Mathematics of Operations Research*, 26(2):248–264, 2001.

Patrick Behr, Andre Guettler, and Felix Miebs. On portfolio optimization: Imposing the right constraints. *Journal of Banking & Finance*, 37(4):1232–1242, 2013. ISSN 0378-4266.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

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*. SIAM, 2025.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, 2011.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optimization Methods and Software*, 27(2):197–219, 2012.

Yair Censor and Arnold Lent. Cyclic subgradient projections. *Mathematical Programming*, 24(1): 233–235, 1982.

Cyrille W Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 49(4):565–571, 2021.

Patrick L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, volume 8 of *Studies in Computational Mathematics*, pp. 115–152. Elsevier, 2001.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations (ICLR)*, 2024.

Qi Deng, Guanghui Lan, and Zhenwei Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *INFORMS Journal on Computing*, 2026.

Nikhil Devanathan and Stephen Boyd. Polyak minorant method for convex optimization. *Journal of Optimization Theory and Applications*, 203(3):2263–2282, 2024.

Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

Michael C. Ferris and S. K. Zavriev. The linear convergence of a successive linear programming algorithm. Technical Report 96-12, Computer Sciences Department, University of Wisconsin–Madison, 1996. URL `https://pages.cs.wisc.edu/~ferris/techreports/96-12.pdf`.

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. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer Berlin, Heidelberg, 1993. doi: 10.1007/978-3-642-78240-4.

Kaja Gruntkowska and Peter Richtárik. Non-Euclidean Broximal Point Method: A blueprint for geometry-aware optimization. *arXiv preprint arXiv:2510.00823*, 2025.

Kaja Gruntkowska, Hanmin Li, Aadi Rane, and Peter Richtárik. The Ball-Proximal (Broximal) Point Method: a new algorithm, convergence theory, and applications. *arXiv preprint arXiv:2502.02002*, 2025.

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

Peter Hall and Brett Presnell. Density estimation under constraints. *Journal of Computational and Graphical Statistics*, 8(2):259–277, 1999. ISSN 10618600.

Richard J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795 – 800, 1985. doi: 10.1214/aos/1176349557.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American symposium on theoretical informatics*, pp. 306–316. Springer, 2008.

Elad Hazan. A linearly convergent projection-free algorithm for smooth convex sets. *arXiv preprint arXiv:2608.04321*, 2026.

Elad Hazan and Sham M. Kakade. Revisiting the Polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

Chang He, Wenzhi Gao, Bo Jiang, Madeleine Udell, and Shuzhong Zhang. New results on the Polyak stepsize: Tight convergence analysis and universal function classes. *arXiv preprint arXiv:2512.06231*, 2025.

Masaru Ito, Zhaosong Lu, and Chuan He. A parameter-free conditional gradient method for composite minimization under hölder condition. *Journal of Machine Learning Research*, 24(166):1–34, 2023.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 427–435, 2013.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. ISSN 0890-5401.

Krzysztof C Kiwiel. The efficiency of subgradient projection methods for convex optimization, part i: General level methods. *SIAM Journal on Control and Optimization*, 34(2):660–676, 1996.

Krzysztof C Kiwiel, Torbjörn Larsson, and Per Olov Lindberg. The efficiency of ballstep subgradient level methods for convex optimization. *Mathematics of Operations Research*, 24(1):237–254, 1999.

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. *arXiv preprint arXiv:1309.5550*, 2013.

Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1):1–45, 2015.

Claude Lemaréchal, Arkadi Nemirovski, and Yurii Nesterov. New variants of bundle methods. *Mathematical Programming*, 69(1):111–147, 1995.

Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *Mathematical Programming*, pp. 219–256, 2026.

David Martínez-Rubio and Sebastian Pokutta. Beyond short steps in Frank–Wolfe algorithms. In *International Conference on Learning Representations (ICLR)*, pp. 157012–157025, 2026.

John Mattingley and Stephen Boyd. Real-time convex optimization in signal processing. *IEEE Signal Processing Magazine*, 27(3):50–61, 2010.

Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171:311–330, 2018.

Francesco Orabona and Ryan D'Orazio. New perspectives on the Polyak stepsize: Surrogate functions and negative results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning (ICML)*, 2025.

Sebastian Pokutta, Christoph Spiegel, and Max Zimmer. Deep neural network training with Frank–Wolfe. *arXiv preprint arXiv:2010.07243*, 2020.

Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From Muon to Gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *International Conference on Machine Learning (ICML)*, 2026.

Peter Richtárik, Kaja Gruntkowska, and Hanmin Li. Local LMO: Constrained gradient optimization via a local linear minimization oracle. *arXiv preprint arXiv:2605.08850*, 2026.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1997.

Sara Sangalli, Ertunc Erdil, Andreas Hötker, Olivio Donati, and Ender Konukoglu. Constrained optimization to train neural networks on critical and under-represented classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mikhail V. Solodov and Benar Fux Svaiter. A new projection method for variational inequality problems. *SIAM Journal on Control and Optimization*, 37(3):765–776, 1999.

Philip Wolfe. Convergence theory in nonlinear programming. *Integer and nonlinear programming*, 3, 1970.

Qisong Yang, Thiago D. Simão, Simon H. Tindemans, and Matthijs T. J. Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887, 2023.

Yijun Zhong and Yi Shen. Stability of constrained optimization models for structured signal recovery. *arXiv preprint arXiv:2601.04849*, 2026.

CONTENTS

1 Introduction **1**

1.1 A family of projection methods 2

1.2 Contributions 3

2 Related Literature **3**

3 Theory **4**

3.1 Local LMO is a projection method 5

3.2 A universal rate for deep cuts 6

3.3 A universal rate for Local LMO 7

4 Experiments **8**

5 Conclusion **9**

A Proof of Theorem 2: Local LMO is a projection method **14**

A.1 Levels and depths 15

A.2 Preliminary lemmas 15

A.3 Localization and Fejér monotonicity 16

A.4 Three thresholds 18

A.5 Identification, step length, and the proof of Theorem 2 19

B Proof of Theorem 3: a universal rate on the deep-cut band **23**

B.1 The Polyak cut turns a Bregman divergence into a function gap 24

B.2 A universal rate for any method that cuts at the Polyak level 24

B.3 Proof of Theorem 3 25

C Proof of Theorem 4: a universal rate for Local LMO **25**

C.1 Two consequences of Hölder continuity 26

C.2 The local Frank–Wolfe gap and the overshoot 27

C.3 The sharpened Fejér budget 28

C.4 Proof of Theorem 4 29

A PROOF OF THEOREM 2: Local LMO IS A PROJECTION METHOD

Proof outline. Theorem 2 establishes that the Local LMO step is the projection of x onto $\mathcal{X} \cap \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$, that its length is exactly t, and that the squared distance to any minimizer decreases by at least t^2.

Our proof of this result takes the following form.

1. Lemmas 5, 7 and 8 establish basic geometric properties that are used throughout this and the following sections.

2. Theorem 9 shows that the set of minimizers, $\mathcal{X}_\star$, lies in every sufficiently shallow half-space $\mathcal{H} = \mathcal{H}(\nabla f(x), \alpha)$ with $\alpha \geq \alpha_\star(x)$. From this, we establish Fejér monotonicity, a key ingredient for the remaining proof steps, for the whole family of projection methods of the form (7) in Theorem 10.
3. Theorem 9 and Lemmas 11 and 12 show that $\alpha_\star(x) \leq \alpha_{\mathrm{cvx}}(x) \leq \alpha_{\mathrm{proj}}(x,t) \leq \alpha_{\mathrm{lmo}}(x,t)$ for admissible t, with $\alpha_{\mathrm{cvx}} = \alpha_{\mathrm{proj}}$ when $t = t(x)$.
4. Theorem 13 shows that the linear minimizer over $\mathcal{X} \cap \mathcal{B}(x,t)$ lies on the sphere $\|z - x\| = t$, and that it satisfies the variational characterization of the projection of x onto $\mathcal{X} \cap \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$; that the distance from x to this set equals t is then a consequence.

A.1 Levels and depths

The main text indexes a cut by its depth, the distance from the current iterate to the cut, as in (5). The proofs below are more naturally written in terms of its level, the value taken by the linear functional $\langle g, \cdot \rangle$ on the cut. For $g \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ we write

$$\mathcal{H}(g, \alpha) \;\coloneqq\; \left\{ z \in \mathbb{R}^d \;:\; \langle g, z \rangle \leq \alpha \right\}, \tag{31}$$

so that, for $g \neq 0$ and $\alpha \leq \langle g, x \rangle$, the two descriptions are interchangeable through

$$\mathcal{H}_x(g, s) \;=\; \mathcal{H}\left(g,\; \langle g, x \rangle - s\|g\|\right), \qquad \text{equivalently} \qquad \mathcal{H}(g, \alpha) \;=\; \mathcal{H}_x\left(g,\; \tfrac{\langle g, x \rangle - \alpha}{\|g\|}\right). \tag{32}$$

On that range the correspondence $\alpha \leftrightarrow s = (\langle g, x \rangle - \alpha)/\|g\|$ is an order-reversing affine bijection onto the depths $s \geq 0$: a larger level is a shallower cut and a larger half-space. Four levels occur below, namely $\alpha_\star(x)$ of (34), $\alpha_{\mathrm{cvx}}(x)$ of (36), $\alpha_{\mathrm{proj}}(x,t)$ of (41) and $\alpha_{\mathrm{lmo}}(x,t)$ of (43). For $x \in \mathcal{X} \setminus \mathcal{X}_\star$ they are the levels of the cuts at depths

$$\alpha_\star(x) \leftrightarrow s_\star(x), \qquad \alpha_{\mathrm{cvx}}(x) \leftrightarrow t(x), \qquad \alpha_{\mathrm{proj}}(x,t) \leftrightarrow t, \qquad \alpha_{\mathrm{lmo}}(x,t) \leftrightarrow s_{\mathrm{lmo}}(x,t), \tag{33}$$

with $s_\star$ as in (9) and s_{lmo} as in (13); the last of these is the identity $s_{\mathrm{lmo}}(x,t) = \big(\langle \nabla f(x), x \rangle - \alpha_{\mathrm{lmo}}(x,t)\big)/\|\nabla f(x)\|$, which holds because $\alpha_{\mathrm{lmo}}(x,t) = \langle \nabla f(x), x_{\mathrm{lmo}} \rangle$. Consequently the level chain $\alpha_\star(x) \leq \alpha_{\mathrm{cvx}}(x) \leq \alpha_{\mathrm{proj}}(x,t) \leq \alpha_{\mathrm{lmo}}(x,t)$ established in Theorem 9 and Lemmas 11 and 12 is the same as the chain of depths $s_{\mathrm{lmo}}(x,t) \leq t \leq t(x) \leq s_\star(x)$ asserted in part 1 of Theorem 2. In the same way, a cut localizes $\mathcal{X}_\star$ if and only if $s \leq s_\star(x)$, equivalently $\alpha \geq \alpha_\star(x)$; and it lies in the deep-cut band, $t(x) \leq s \leq s_\star(x)$, if and only if $\alpha_\star(x) \leq \alpha \leq \alpha_{\mathrm{cvx}}(x)$.

A.2 Preliminary lemmas

Lemma 5. *If $x \in \mathcal{X} \setminus \mathcal{X}_\star$, then $\nabla f(x) \neq 0$.*

Proof. We show the contrapositive. Suppose $x \in \mathcal{X}$ and $\nabla f(x) = 0$. Then, by convexity and differentiability of f (Assumption 1), we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle = f(x), \qquad \forall y \in \mathbb{R}^d.$$

Thus, x is a global minimizer of f, and, in particular, a minimizer over the feasible set $\mathcal{X}$. Hence, $x \in \mathcal{X}_\star$. □

Lemma 6. *The gradient of f is constant on the solution set: $\nabla f(u) = \nabla f(v)$ for all $u, v \in \mathcal{X}_\star$. In particular, $G_\star \coloneqq \|\nabla f(x_\star)\|$ does not depend on the choice of $x_\star \in \mathcal{X}_\star$.*

Proof. Let $u, v \in \mathcal{X}_\star$. First-order optimality of v for f over $\mathcal{X}$ gives $\langle \nabla f(v), u - v \rangle \geq 0$, while $f(u) = f(v) = f_\star$, so

$$D_f(u, v) \;=\; f(u) - f(v) - \langle \nabla f(v), u - v \rangle \;=\; -\langle \nabla f(v), u - v \rangle \;\leq\; 0.$$

Convexity of f gives $D_f(u, v) \geq 0$, hence $D_f(u, v) = 0$, that is $\langle \nabla f(v), u - v \rangle = 0$. Therefore, for every $w \in \mathbb{R}^d$, convexity at v yields

$$f(w) \;\geq\; f(v) + \langle \nabla f(v), w - v \rangle \;=\; f(u) + \langle \nabla f(v), w - u \rangle,$$

the last equality because $f(v) = f(u)$ and $\langle \nabla f(v), u - v \rangle = 0$. This is the subgradient inequality for $\nabla f(v)$ at u, so $\nabla f(v) \in \partial f(u) = \{\nabla f(u)\}$, the last set equality because f is differentiable (Rockafellar, 1997, Theorem 25.1). □

Lemma 7. *For any three points $x_+, x, z \in \mathbb{R}^d$,*

$$\|x_+ - z\|^2 = \|x - z\|^2 - 2\langle x - x_+, x_+ - z\rangle - \|x - x_+\|^2.$$

Proof. We write $x - z = (x - x_+) + (x_+ - z)$. Hence,

$$\|x - z\|^2 = \|(x - x_+) + (x_+ - z)\|^2 = \|x - x_+\|^2 + \|x_+ - z\|^2 + 2\langle x - x_+, x_+ - z\rangle.$$

Rearranging this identity yields

$$\|x_+ - z\|^2 = \|x - z\|^2 - 2\langle x - x_+, x_+ - z\rangle - \|x - x_+\|^2.$$

□

Lemma 8. *Let $x \in \mathbb{R}^d$ and $t > 0$, and define the closed Euclidean ball*

$$\mathcal{B}(x,t) := \left\{z \in \mathbb{R}^d \ : \ \|z - x\| \le t\right\}.$$

Then, for any $y \in \mathbb{R}^d$, the normal cone of $\mathcal{B}(x,t)$ at y is given by

$$N_{\mathcal{B}(x,t)}(y) = \begin{cases} \emptyset, & y \notin \mathcal{B}(x,t), \\ \{0\}, & \|y - x\| < t, \\ \{\lambda(y - x) \ : \ \lambda \ge 0\}, & \|y - x\| = t. \end{cases}$$

Proof. Recall that, for a convex set $\mathcal{C}$ and $y \in \mathcal{C}$, $N_{\mathcal{C}}(y) = \left\{v \in \mathbb{R}^d \ : \ \langle v, z - y\rangle \le 0 \ \forall z \in \mathcal{C}\right\}$, while $N_{\mathcal{C}}(y) = \emptyset$ for $y \notin \mathcal{C}$. This settles the case $y \notin \mathcal{B}(x,t)$.

For the remainder of this proof, we write $\mathcal{B} := \mathcal{B}(x,t)$.

Case $\|y - x\| < t$. Clearly $0 \in N_{\mathcal{B}}(y)$. Let $v \in N_{\mathcal{B}}(y)$ and suppose $v \neq 0$. Set $\varepsilon := t - \|y - x\| > 0$ and $z := y + \varepsilon v/\|v\|$. Then $\|z - x\| \le \|y - x\| + \varepsilon = t$, so $z \in \mathcal{B}$, but $\langle v, z - y\rangle = \varepsilon\|v\| > 0$, a contradiction. Hence $N_{\mathcal{B}}(y) = \{0\}$.

Case $\|y - x\| = t$. Let $\lambda \ge 0$ and $z \in \mathcal{B}$. By Cauchy–Schwarz,

$$\begin{aligned} \langle \lambda(y - x), z - y\rangle &= \lambda\left(\langle y - x, z - x\rangle - \|y - x\|^2\right) \\ &\le \lambda\left(\|y - x\|\|z - x\| - t^2\right) \\ &\le \lambda\left(t^2 - t^2\right) = 0, \end{aligned}$$

so $\lambda(y - x) \in N_{\mathcal{B}}(y)$.

Next, let $v \in N_{\mathcal{B}}(y)$, set $u := (y - x)/t$ (a unit vector), and decompose $v = \mu u + w$ with $\mu := \langle v, u\rangle$ and $\langle w, u\rangle = 0$. Choosing $z = x \in \mathcal{B}$ gives $0 \ge \langle v, x - y\rangle = -t\mu$, so $\mu \ge 0$. Suppose $w \neq 0$, let $\hat{w} := w/\|w\|$, and for $\beta \in (0, \pi/2]$ consider $z_\beta := x + t(\cos\beta\, u + \sin\beta\, \hat{w})$. Since $u \perp \hat{w}$ are unit vectors, $\|z_\beta - x\| = t$, so $z_\beta \in \mathcal{B}$. Using $z_\beta - y = t((\cos\beta - 1)u + \sin\beta\, \hat{w})$, $1 - \cos\beta \le \beta^2/2$ and $\sin\beta \ge 2\beta/\pi$,

$$\langle v, z_\beta - y\rangle = t(\mu(\cos\beta - 1) + \|w\|\sin\beta) \ge t\beta\left(\tfrac{2}{\pi}\|w\| - \tfrac{\mu}{2}\beta\right),$$

which is positive for all sufficiently small $\beta > 0$, contradicting $v \in N_{\mathcal{B}}(y)$. Therefore $w = 0$ and $v = \mu u = (\mu/t)(y - x)$ with $\mu/t \ge 0$. □

A.3 Localization and Fejér monotonicity

The following two results underlie parts 1 and 4 of Theorem 2, respectively. Theorem 10 further underlies the Fejér arguments of Sections B and C.

Theorem 9. *Fix a point $x \in \mathbb{R}^d$.*

1. *The quantity*

$$\alpha_\star(x) := \sup_{x_\star \in \mathcal{X}_\star} \langle \nabla f(x), x_\star\rangle \tag{34}$$

is finite. In particular,

$$\alpha_\star(x) \le \alpha_{\mathrm{cvx}}(x), \tag{35}$$

where

$$\alpha_{\mathrm{cvx}}(x) := \langle \nabla f(x), x\rangle - (f(x) - f_\star) < +\infty. \tag{36}$$

2. *For $\alpha \in \mathbb{R}$, the following statements are equivalent:*
   i. $\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha)$,
   ii. $\langle \nabla f(x), x_\star \rangle \leq \alpha$ *for every* $x_\star \in \mathcal{X}_\star$,
   iii. $\alpha_\star(x) \leq \alpha$.
3. *The following inclusions hold:*

$$\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha_\star(x)) \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{cvx}}(x)). \tag{37}$$

Proof. 1. Let $x_\star \in \mathcal{X}_\star$ be arbitrary. Since f is convex and differentiable on $\mathbb{R}^d$ (Assumption 1), the first-order convexity inequality gives

$$f_\star = f(x_\star) \geq f(x) + \langle \nabla f(x), x_\star - x \rangle .$$

Rearranging, we obtain

$$\langle \nabla f(x), x_\star \rangle \leq \langle \nabla f(x), x \rangle - (f(x) - f_\star) = \alpha_{\mathrm{cvx}}(x) < +\infty.$$

Since this inequality holds for every $x_\star \in \mathcal{X}_\star$, we may take the supremum over $x_\star \in \mathcal{X}_\star$ and get

$$\alpha_\star(x) = \sup_{x_\star \in \mathcal{X}_\star} \langle \nabla f(x), x_\star \rangle \leq \alpha_{\mathrm{cvx}}(x) < +\infty.$$

Moreover, since $\mathcal{X}_\star \neq \emptyset$ (Assumption 1) the supremum is over a non-empty set, hence $\alpha_\star(x) > -\infty$. This proves (35).

2. Let $g = \nabla f(x)$. By definition of $\mathcal{H}(g, \alpha)$, for any $x_\star \in \mathcal{X}_\star$,

$$x_\star \in \mathcal{H}(g, \alpha) \qquad \Longleftrightarrow \qquad \langle g, x_\star \rangle \leq \alpha.$$

Therefore,

$$\mathcal{X}_\star \subseteq \mathcal{H}(g, \alpha) \qquad \Longleftrightarrow \qquad \langle g, x_\star \rangle \leq \alpha \quad \forall x_\star \in \mathcal{X}_\star.$$

This proves the equivalence of i. and ii.
Next, statement ii. is equivalent to saying that α is an upper bound on the set

$$\{ \langle g, x_\star \rangle \ : \ x_\star \in \mathcal{X}_\star \} .$$

By the definition (34) of $\alpha_\star$, this is equivalent to

$$\sup_{x_\star \in \mathcal{X}_\star} \langle g, x_\star \rangle \leq \alpha,$$

that is, $\alpha_\star(x) \leq \alpha$. This proves the equivalence of ii. and iii.

3. This follows immediately by combining part 2 and (35), and using the definition (31).

□

Theorem 10. *Fix $x_0 \in \mathcal{X}$, and define*

$$x_{k+1} \coloneqq \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}(\nabla f(x_k), \alpha_k)}(x_k), \qquad k = 0, 1, \ldots,$$

where $\alpha_k \geq \alpha_\star(x_k)$. Then, for all $k \geq 0$, we have

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \|x_k - x_{k+1}\|^2 \qquad \forall x_\star \in \mathcal{X}_\star, \tag{38}$$

and therefore, for all $K \geq 1$, we have

$$\min_{0 \leq k \leq K-1} \|x_k - x_{k+1}\|^2 \leq \frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}. \tag{39}$$

Proof. By Lemma 7, we know that

$$\|x_{k+1} - x_\star\|^2 = \|x_k - x_\star\|^2 - 2 \langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle - \|x_k - x_{k+1}\|^2$$

for all $x_\star \in \mathcal{X}_\star$. Hence, we only need to show that

$$\langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle \geq 0. \tag{40}$$

Let $g_k = \nabla f(x_k)$. Since $\alpha_k \geq \alpha_\star(x_k)$, Theorem 9 implies that $\mathcal{X}_\star \subseteq \mathcal{H}(g_k, \alpha_k)$. Since also $\mathcal{X}_\star \subseteq \mathcal{X}$, we have

$$x_\star \in \mathcal{X} \cap \mathcal{H}(g_k, \alpha_k), \qquad \forall x_\star \in \mathcal{X}_\star.$$

Because x_{k+1} is the Euclidean projection of x_k onto the non-empty closed convex set $\mathcal{X} \cap \mathcal{H}(g_k, \alpha_k)$, the standard projection optimality condition yields

$$\langle x_k - x_{k+1}, y - x_{k+1} \rangle \leq 0, \qquad \forall y \in \mathcal{X} \cap \mathcal{H}(g_k, \alpha_k).$$

Since $\mathcal{X}_\star \subseteq \mathcal{X} \cap \mathcal{H}(g_k, \alpha_k)$, the above inequality holds for all $y = x_\star \in \mathcal{X}_\star$, establishing (40).

Finally, (39) follows by summing up inequalities (38) for $k = 0, \ldots, K-1$. Indeed, the inequalities

$$\begin{aligned} \|x_1 - x_\star\|^2 &\leq \|x_0 - x_\star\|^2 - \|x_0 - x_1\|^2 \\ \|x_2 - x_\star\|^2 &\leq \|x_1 - x_\star\|^2 - \|x_1 - x_2\|^2 \\ &\vdots \\ \|x_K - x_\star\|^2 &\leq \|x_{K-1} - x_\star\|^2 - \|x_{K-1} - x_K\|^2 \end{aligned}$$

add up to

$$\|x_K - x_\star\|^2 \leq \|x_0 - x_\star\|^2 - \sum_{k=0}^{K-1} \|x_k - x_{k+1}\|^2,$$

and after rearranging and dividing by K, we get

$$\min_{0 \leq k \leq K-1} \|x_k - x_{k+1}\|^2 \leq \frac{1}{K} \sum_{k=0}^{K-1} \|x_k - x_{k+1}\|^2 \leq \frac{\|x_0 - x_\star\|^2 - \|x_K - x_\star\|^2}{K} \leq \frac{\|x_0 - x_\star\|^2}{K}.$$

It only remains to choose

$$x_\star = \operatorname*{argmin}_{z \in \mathcal{X}_\star} \|x_0 - z\|,$$

so that $\|x_0 - x_\star\| = \operatorname{dist}(x_0, \mathcal{X}_\star)$. □

A.4 THREE THRESHOLDS

For $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$, we define

$$\alpha_{\mathrm{proj}}(x, t) \coloneqq \langle \nabla f(x), x \rangle - t \|\nabla f(x)\|. \tag{41}$$

Lemma 11. *Fix $x \in \mathbb{R}^d$ and assume $\nabla f(x) \neq 0$.*

1. *If $t \leq t(x)$, then*
$$\alpha_{\mathrm{cvx}}(x) \leq \alpha_{\mathrm{proj}}(x, t).$$
2. *If $t = t(x)$, then*
$$\alpha_{\mathrm{cvx}}(x) = \alpha_{\mathrm{proj}}(x, t).$$

Consequently, if $t \leq t(x)$, then

$$\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{cvx}}(x)) \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{proj}}(x, t)). \tag{42}$$

Proof. By definition, $\alpha_{\mathrm{cvx}}(x) = \langle \nabla f(x), x \rangle - (f(x) - f_\star)$, while $\alpha_{\mathrm{proj}}(x, t) = \langle \nabla f(x), x \rangle - t\|\nabla f(x)\|$. Hence

$$\alpha_{\mathrm{proj}}(x, t) - \alpha_{\mathrm{cvx}}(x) = (f(x) - f_\star) - t\|\nabla f(x)\|.$$

Using the definition (4) of $t(x)$, we can rewrite this as

$$\alpha_{\mathrm{proj}}(x, t) - \alpha_{\mathrm{cvx}}(x) = \|\nabla f(x)\| \left(t(x) - t\right).$$

Therefore, if $t \leq t(x)$, then $\alpha_{\mathrm{proj}}(x, t) - \alpha_{\mathrm{cvx}}(x) \geq 0$, which proves $\alpha_{\mathrm{cvx}}(x) \leq \alpha_{\mathrm{proj}}(x, t)$. If $t = t(x)$, then $\alpha_{\mathrm{proj}}(x, t) - \alpha_{\mathrm{cvx}}(x) = 0$, and thus $\alpha_{\mathrm{cvx}}(x) = \alpha_{\mathrm{proj}}(x, t)$. The first inclusion in (42) is (35) of Theorem 9 together with part 2 of that theorem; the second is $\alpha_{\mathrm{cvx}}(x) \leq \alpha_{\mathrm{proj}}(x, t)$ and the monotonicity of $\mathcal{H}(\nabla f(x), \cdot)$. □

Lemma 12. *Fix $x \in \mathcal{X} \setminus \mathcal{X}_\star$. For $t > 0$, define*

$$\alpha_{\mathrm{lmo}}(x,t) \coloneqq \min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle . \tag{43}$$

Then

$$\alpha_{\mathrm{proj}}(x,t) \le \alpha_{\mathrm{lmo}}(x,t). \tag{44}$$

Consequently, if $0 < t \le t(x)$, then

$$\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{proj}}(x,t)) \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t)). \tag{45}$$

Proof. Let $g = \nabla f(x)$, $\alpha_{\mathrm{lmo}} = \alpha_{\mathrm{lmo}}(x,t)$ and $\alpha_{\mathrm{proj}} = \alpha_{\mathrm{proj}}(x,t)$.

Step 1: Non-zero gradient. Since $x \notin \mathcal{X}_\star$, we have $g \neq 0$ (Lemma 5).

Step 2: Inequality between the thresholds. Fix any $z \in \mathcal{X} \cap \mathcal{B}(x,t)$. Since $z \in \mathcal{B}(x,t)$, we have $\|z - x\| \le t$. Therefore, by the Cauchy–Schwarz inequality,

$$\langle g, z \rangle = \langle g, x \rangle + \langle g, z - x \rangle \ge \langle g, x \rangle - \|g\| \, \|z - x\| \ge \langle g, x \rangle - t\|g\| = \alpha_{\mathrm{proj}}.$$

Since this holds for every $z \in \mathcal{X} \cap \mathcal{B}(x,t)$, taking the minimum over $z \in \mathcal{X} \cap \mathcal{B}(x,t)$ yields

$$\alpha_{\mathrm{lmo}} = \min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, z \rangle \ge \alpha_{\mathrm{proj}}.$$

This proves (44).

Step 3: Relationship between the half-spaces. Now assume in addition that $0 < t \le t(x)$. By (42) of Lemma 11, we have

$$\mathcal{X}_\star \subseteq \mathcal{H}(g, \alpha_{\mathrm{proj}}).$$

Moreover, the inclusion

$$\mathcal{H}(g, \alpha_{\mathrm{proj}}) \subseteq \mathcal{H}(g, \alpha_{\mathrm{lmo}})$$

follows from (44). Combining the last two inclusions establishes (45). □

A.5 Identification, step length, and the proof of Theorem 2

Theorem 13. *Fix $x \in \mathcal{X} \setminus \mathcal{X}_\star$, choose*

$$0 < t \le t(x) \coloneqq \frac{f(x) - f_\star}{\|\nabla f(x)\|}, \tag{46}$$

and let

$$x_{\mathrm{lmo}} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle , \tag{47}$$

with

$$\alpha_{\mathrm{lmo}}(x,t) \coloneqq \min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle = \langle \nabla f(x), x_{\mathrm{lmo}} \rangle . \tag{48}$$

Then,

1. *x_{lmo} lies on the boundary of the ball $\mathcal{B}(x,t)$, that is,*

$$\|x_{\mathrm{lmo}} - x\| = t; \tag{49}$$

2. *x_{lmo} is the Euclidean projection of x onto the intersection of $\mathcal{X}$ with the half-space*

$$\mathcal{H}_{\mathrm{lmo}} \coloneqq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t)),$$

that is,

$$x_{\mathrm{lmo}} = \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_{\mathrm{lmo}}}(x). \tag{50}$$

Proof. Let $g \coloneqq \nabla f(x)$.

Step 1: Non-zero gradient. Since $x \notin \mathcal{X}_\star$, we have $\nabla f(x) \neq 0$ (Lemma 5).

Step 2: Non-zero step. We now show that

$$x_{\mathrm{lmo}} \neq x, \tag{51}$$

which will be useful later.

Assume, for the sake of contradiction, that $x_{\mathrm{lmo}} = x$. Since x_{lmo} solves (47), this implies that x minimizes the linear functional

$$z \mapsto \langle \nabla f(x), z \rangle$$

over $\mathcal{X} \cap \mathcal{B}(x,t)$. Hence,

$$\langle \nabla f(x), z - x \rangle \geq 0, \qquad \forall z \in \mathcal{X} \cap \mathcal{B}(x,t). \tag{52}$$

Now fix any $x_\star \in \mathcal{X}_\star$. Since f is convex and differentiable (Assumption 1), we have

$$f_\star = f(x_\star) \geq f(x) + \langle \nabla f(x), x_\star - x \rangle . \tag{53}$$

Since $x \notin \mathcal{X}_\star$, we have $f(x) > f_\star$, and therefore

$$\langle \nabla f(x), x - x_\star \rangle \overset{(53)}{\geq} f(x) - f_\star > 0. \tag{54}$$

Because the radius t satisfies (46), and

$$f(x) - f_\star \overset{(53)}{\leq} \langle \nabla f(x), x - x_\star \rangle \leq \|\nabla f(x)\| \, \|x - x_\star\|,$$

it follows that $0 < t \leq \|x - x_\star\|$. The scalar $\lambda \coloneqq \frac{t}{\|x - x_\star\|}$ thus belongs to $(0, 1]$. Define

$$z_t \coloneqq (1-\lambda)x + \lambda x_\star = x - \frac{t}{\|x - x_\star\|}(x - x_\star). \tag{55}$$

Since $x, x_\star \in \mathcal{X}$ and $\mathcal{X}$ is convex (Assumption 1), we have $z_t \in \mathcal{X}$ also. Therefore,

$$\|z_t - x\| \overset{(55)}{=} \frac{t}{\|x - x_\star\|}\|x - x_\star\| = t.$$

Hence, $z_t \in \mathcal{B}(x,t)$, which implies that $z_t \in \mathcal{X} \cap \mathcal{B}(x,t)$. Applying (52) with $z = z_t$, we obtain

$$\langle \nabla f(x), z_t - x \rangle \geq 0.$$

On the other hand,

$$\langle \nabla f(x), z_t - x \rangle \overset{(55)}{=} -\frac{t}{\|x - x_\star\|} \langle \nabla f(x), x - x_\star \rangle .$$

By (54), the right-hand side is strictly negative, and hence

$$\langle \nabla f(x), z_t - x \rangle < 0,$$

which contradicts inequality (52). We therefore conclude that $x_{\mathrm{lmo}} \neq x$.

Step 3: An upper bound on the radius. Fix any $x_\star \in \mathcal{X}_\star$. Since f is convex and differentiable (Assumption 1),

$$f_\star = f(x_\star) \geq f(x) + \langle \nabla f(x), x_\star - x \rangle . \tag{56}$$

Combining this with (46), we get

$$0 < t \overset{(46)}{\leq} t(x) \overset{(4)}{=} \frac{f(x) - f_\star}{\|\nabla f(x)\|} \overset{(56)}{\leq} \frac{\langle \nabla f(x), x - x_\star \rangle}{\|\nabla f(x)\|}. \tag{57}$$

Step 4: Optimality conditions for the local linear minimization problem. Since x_{lmo} solves the optimization problem

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z \rangle ,$$

the first-order optimality condition (Rockafellar, 1997, Theorem 27.4) gives $0 \in \nabla f(x) + N_{\mathcal{X} \cap \mathcal{B}(x,t)}(x_{\mathrm{lmo}})$, where $N_{\mathcal{C}}(z)$ denotes the normal cone of the set $\mathcal{C}$ at point z. The sum rule $N_{\mathcal{X} \cap \mathcal{B}(x,t)} = N_{\mathcal{X}} + N_{\mathcal{B}(x,t)}$ applies here as both sets are convex and their relative interiors meet

(Rockafellar, 1997, Corollary 23.8.1). Indeed, $\operatorname{ri}\mathcal{B}(x,t) = \operatorname{int}\mathcal{B}(x,t) = \{z \;:\; \|z-x\| < t\}$ is an open set containing x. Since $\mathcal{X}$ is non-empty and convex, $x \in \mathcal{X} \subseteq \operatorname{cl}\mathcal{X} = \operatorname{cl}(\operatorname{ri}\mathcal{X})$ (Rockafellar, 1997, Theorem 6.3), so this open neighborhood of x contains a point of $\operatorname{ri}\mathcal{X}$, i.e. $\operatorname{ri}\mathcal{X} \cap \operatorname{ri}\mathcal{B}(x,t) \neq \emptyset$.

Lemma 8 says that

$$N_{\mathcal{B}(x,t)}(y) = \begin{cases} \emptyset, & y \notin \mathcal{B}(x,t), \\ \{0\}, & \|y-x\| < t, \\ \{\lambda'(y-x) \;:\; \lambda' \geq 0\}, & \|y-x\| = t. \end{cases} \tag{58}$$

Since $x_{\text{lmo}} \in \mathcal{B}(x,t)$, we conclude that

$$N_{\mathcal{B}(x,t)}(x_{\text{lmo}}) = \begin{cases} \{0\}, & \|x_{\text{lmo}}-x\| < t, \\ \{\lambda'(x_{\text{lmo}}-x) \;:\; \lambda' \geq 0\}, & \|x_{\text{lmo}}-x\| = t. \end{cases} \tag{59}$$

Hence, there exist $n \in N_{\mathcal{X}}(x_{\text{lmo}})$ and $\lambda \geq 0$ such that

$$\nabla f(x) + n + \lambda(x_{\text{lmo}} - x) = 0. \tag{60}$$

Step 5: Moving to the boundary of the ball. Claim 1 has been established by Richtárik et al. (2026, Theorem 3.1(ii)). We include the proof as we re-use one of the intermediate results below.

We already know that $x_{\text{lmo}} \in \mathcal{B}(x,t)$, and therefore

$$\|x_{\text{lmo}} - x\| \leq t. \tag{61}$$

It remains to prove the reverse inequality. We distinguish two cases.

• **Case 1:** $\lambda > 0$**.** By the construction of (60), $\lambda(x_{\text{lmo}} - x) \in N_{\mathcal{B}(x,t)}(x_{\text{lmo}})$. Since $x_{\text{lmo}} \neq x$ and $\lambda > 0$, characterization (58) of the normal cone $N_{\mathcal{B}(x,t)}(x_{\text{lmo}})$ implies that

$$\|x_{\text{lmo}} - x\| = t.$$

• **Case 2:** $\lambda = 0$**.** In this case, (60) becomes $\nabla f(x) + n = 0$, i.e.,

$$-\nabla f(x) \in N_{\mathcal{X}}(x_{\text{lmo}}).$$

By the definition of the normal cone, this implies $\langle -\nabla f(x), z - x_{\text{lmo}} \rangle \leq 0$ for all $z \in \mathcal{X}$. Equivalently,

$$\langle \nabla f(x), x_{\text{lmo}} - z \rangle \leq 0, \qquad \forall z \in \mathcal{X}.$$

In particular, for any $x_\star \in \mathcal{X}_\star \subseteq \mathcal{X}$,

$$\langle \nabla f(x), x_{\text{lmo}} - x_\star \rangle \leq 0. \tag{62}$$

Using the decomposition

$$x_{\text{lmo}} - x_\star = (x - x_\star) + (x_{\text{lmo}} - x),$$

we obtain

$$\langle \nabla f(x), x_{\text{lmo}} - x_\star \rangle = \langle \nabla f(x), x - x_\star \rangle + \langle \nabla f(x), x_{\text{lmo}} - x \rangle. \tag{63}$$

By the Cauchy–Schwarz inequality,

$$\langle \nabla f(x), x_{\text{lmo}} - x \rangle \geq -\|\nabla f(x)\| \|x_{\text{lmo}} - x\| \overset{(61)}{\geq} -\|\nabla f(x)\| t. \tag{64}$$

By combining this with (63), we get

$$\langle \nabla f(x), x_{\text{lmo}} - x_\star \rangle \geq \langle \nabla f(x), x - x_\star \rangle - t\|\nabla f(x)\|.$$

Using (57), we conclude that

$$\langle \nabla f(x), x_{\text{lmo}} - x_\star \rangle \geq 0. \tag{65}$$

Combining (62) and (65), we get

$$\langle \nabla f(x), x_{\text{lmo}} - x_\star \rangle = 0.$$

Hence equality must hold in the preceding chain of inequalities, and in particular in (64). This implies

$$\|x_{\text{lmo}} - x\| = t,$$

which proves (49).

Step 6: Projection characterization. We now prove that

$$x_{\mathrm{lmo}} = \mathrm{Proj}_{\mathcal{X}\cap\mathcal{H}(\nabla f(x),\alpha_{\mathrm{lmo}}(x,t))}(x).$$

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a non-empty closed convex set and let $x \in \mathbb{R}^d$. A point $u \in \mathcal{C}$ satisfies

$$u = \mathrm{Proj}_{\mathcal{C}}(x)$$

if and only if

$$\langle x - u, y - u\rangle \le 0, \qquad \forall y \in \mathcal{C}. \tag{66}$$

Therefore, it suffices to prove that

$$\langle x - x_{\mathrm{lmo}}, y - x_{\mathrm{lmo}}\rangle \le 0, \qquad \forall y \in \mathcal{X} \cap \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t)). \tag{67}$$

We again distinguish the same two cases as above.

• **Case 1:** $\lambda > 0$**.** Fix any $y \in \mathcal{X} \cap \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$. Since $n \in N_{\mathcal{X}}(x_{\mathrm{lmo}})$ and $y \in \mathcal{X}$,

$$\langle n, y - x_{\mathrm{lmo}}\rangle \le 0.$$

Also, since $y \in \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$ and $\alpha_{\mathrm{lmo}}(x,t) = \langle \nabla f(x), x_{\mathrm{lmo}}\rangle$,

$$\langle \nabla f(x), y - x_{\mathrm{lmo}}\rangle \le 0.$$

Taking the inner product of (60) with $y - x_{\mathrm{lmo}}$, we obtain

$$\langle \nabla f(x), y - x_{\mathrm{lmo}}\rangle + \langle n, y - x_{\mathrm{lmo}}\rangle + \lambda \langle x_{\mathrm{lmo}} - x, y - x_{\mathrm{lmo}}\rangle = 0.$$

The first two terms are non-positive, and hence $\lambda \langle x_{\mathrm{lmo}} - x, y - x_{\mathrm{lmo}}\rangle \ge 0$. Since $\lambda > 0$,

$$\langle x - x_{\mathrm{lmo}}, y - x_{\mathrm{lmo}}\rangle \le 0.$$

Thus (67) holds in this case.

• **Case 2:** $\lambda = 0$**.** As proved above, equality must hold in the Cauchy–Schwarz inequality (64), and therefore

$$x - x_{\mathrm{lmo}} = t\,\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Now fix any $y \in \mathcal{X} \cap \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$. Since $y \in \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$,

$$\langle \nabla f(x), y - x_{\mathrm{lmo}}\rangle \le 0.$$

Therefore

$$\langle x - x_{\mathrm{lmo}}, y - x_{\mathrm{lmo}}\rangle = \frac{t}{\|\nabla f(x)\|}\,\langle \nabla f(x), y - x_{\mathrm{lmo}}\rangle \le 0.$$

Thus (67) holds in this case as well. □

We now have all the parts to establish Theorem 2, which we restate here for convenience.

Theorem 2. *Fix $x \in \mathcal{X} \setminus \mathcal{X}_\star$ and an admissible radius $t \in (0, t(x)]$, and let $x_{\mathrm{lmo}} \in \mathrm{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x,t)} \langle \nabla f(x), z\rangle$ be a* Local LMO *step. Then:*

1. *The* Local LMO *depth is admissible and does not exceed the radius, that is,*

$$0 \;<\; s_{\mathrm{lmo}}(x,t) \;\le\; t \;\le\; t(x) \;\le\; s_\star(x),$$

so the half-space $\mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right)$ separates x from, and localizes, the set of minimizers:

$$x \notin \mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right), \qquad \mathcal{X}_\star \subseteq \mathcal{H}_x\left(\nabla f(x), s_{\mathrm{lmo}}(x,t)\right).$$

2. *The projection onto the localized feasible set* (7) *coincides with the* Local LMO *step, that is,*

$$x_{\mathrm{lmo}} = \mathrm{Proj}_{\mathcal{X}\cap\mathcal{H}_x(\nabla f(x),\, s_{\mathrm{lmo}}(x,t))}(x).$$

In particular, $\mathrm{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x,t)} \langle \nabla f(x), z\rangle$ *is a singleton.*

3. *The method takes a step of size t, that is,*

$$\|x_{\mathrm{lmo}} - x\| = \operatorname{dist}\left(x, \mathcal{X} \cap \mathcal{H}_x(\nabla f(x),\, s_{\mathrm{lmo}}(x,t))\right) = t.$$

4. *The squared distance of the new point to any minimizer is reduced by at least t^2, that is,*

$$\|x_{\mathrm{lmo}} - x_\star\|^2 \le \|x - x_\star\|^2 - t^2, \quad \forall x_\star \in \mathcal{X}_\star.$$

The theorem is phrased in depth terms; we argue in levels and translate with (32) and (33), under which $\mathcal{H}_x(\nabla f(x), s_{\mathrm{lmo}}(x,t)) = \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t))$.

Proof. Write $g \coloneqq \nabla f(x)$, which is non-zero by Lemma 5.

1. Theorem 9 and Lemmas 11 and 12 show that $\alpha_\star(x) \le \alpha_{\mathrm{cvx}}(x) \le \alpha_{\mathrm{proj}}(x,t) \le \alpha_{\mathrm{lmo}}(x,t)$, which by (33) is the chain $s_{\mathrm{lmo}}(x,t) \le t \le t(x) \le s_\star(x)$. For the remaining strict inequality $s_{\mathrm{lmo}}(x,t) > 0$, fix any $x_\star \in \mathcal{X}_\star$ and set $\mu \coloneqq t/\|x - x_\star\|$, which lies in $(0,1]$ because $0 < t \le t(x) \le \langle g, x - x_\star\rangle / \|g\| \le \|x - x_\star\|$ by (57) and Cauchy–Schwarz. The point $z \coloneqq (1-\mu)x + \mu x_\star$ lies in $\mathcal{X} \cap \mathcal{B}(x,t)$, so $\alpha_{\mathrm{lmo}}(x,t) \le \langle g, z\rangle = \langle g, x\rangle - \mu\langle g, x - x_\star\rangle$ and therefore $s_{\mathrm{lmo}}(x,t) = (\langle g, x\rangle - \alpha_{\mathrm{lmo}}(x,t))/\|g\| \ge \mu\langle g, x - x_\star\rangle / \|g\| \ge \mu\,(f(x) - f_\star)/\|g\| > 0$; equivalently, $x \notin \mathcal{H}_x(\nabla f(x), s_{\mathrm{lmo}}(x,t))$. The same computation is recorded, in the notation of the local Frank–Wolfe gap, as Lemma 22. Since $\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha_\star(x))$ (Theorem 9) and the half-spaces are increasing in the level α, we have

$$\mathcal{X}_\star \subseteq \mathcal{H}(\nabla f(x), \alpha_\star(x)) \subseteq \mathcal{H}(\nabla f(x), \alpha_{\mathrm{lmo}}(x,t)).$$

2. This is the content of (50), part 2 of Theorem 13.
3. This follows from (49), part 1 of Theorem 13. Further, by part 2 of the same theorem, $x_{\mathrm{lmo}} = \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_{\mathrm{lmo}}}(x)$, so $\operatorname{dist}(x, \mathcal{X} \cap \mathcal{H}_{\mathrm{lmo}}) = \|x - x_{\mathrm{lmo}}\| = t$.
4. Theorem 10 applied with $\alpha = \alpha_{\mathrm{lmo}}(x,t) \ge \alpha_\star(x)$, which holds by part 1, combined with the step length in part 3.
   Since the minimizer is identified with a projection, it is unique: by part 3, every element of $\operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, z\rangle$ lies on the sphere $\|z - x\| = t$, and that argmin is convex, so it is a singleton.

□

The following one-step consequences are used repeatedly in Sections B and C.

Lemma 14. *Let $\{x_k\}$ be generated by* Local LMO *with admissible radii $0 < t_k \le t(x_k)$ from $x_0 \in \mathcal{X}$. Then, for every $k \ge 0$ and every $x_\star \in \mathcal{X}_\star$,*

1. $t_k \le \|x_k - x_\star\|$,
2. $\|x_{k+1} - x_k\| = t_k$,
3. $\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - t_k^2$.

Proof. Let $g_k = \nabla f(x_k)$. Convexity and Cauchy–Schwarz give $f(x_k) - f_\star \le \langle g_k, x_k - x_\star\rangle \le \|g_k\|\|x_k - x_\star\|$, so $t_k \le t(x_k) = (f(x_k) - f_\star)/\|g_k\| \le \|x_k - x_\star\|$, which is part 1. Parts 2 and 3 correspond to parts 3 and 4 of Theorem 2, respectively. □

B Proof of Theorem 3: a universal rate on the deep-cut band

Proof outline. Theorem 3 states that every projection method whose cut lies in the deep-cut band attains the rate $K^{-(1+\vartheta)/2}$, without knowledge of ϑ or L_ϑ. Our argument is based on two observations.

1. Lemma 15 shows that a point x_+ in the Polyak half-space $\mathcal{H}(\nabla f(x), \alpha_{\mathrm{cvx}}(x))$ satisfies $f(x_+) - f_\star \le D_f(x_+, x)$. Cutting at or below the Polyak level converts the Bregman divergence into a function gap.

2. Fejér monotonicity (Theorem 10) makes the squared step lengths summable, so the shortest of K steps is at most $\mathrm{dist}(x_0, \mathcal{X}_\star)/\sqrt{K}$.

Hölder continuity turns the first into $f(x_{k+1}) - f_\star \le \frac{L_\vartheta}{1+\vartheta}\|x_{k+1} - x_k\|^{1+\vartheta}$, and the second then gives the rate. We stress that neither step uses ϑ or L_ϑ, which is why the bound is universal.

Theorems 9 and 10 of Section A are re-used below.

B.1 The Polyak cut turns a Bregman divergence into a function gap

Lemma 15. *Fix $x \in \mathcal{X}$ and let $x_+ \in \mathcal{H}(\nabla f(x), \alpha_{\mathrm{cvx}}(x))$. Then*

$$f(x_+) - f_\star \le D_f(x_+, x). \tag{68}$$

Proof. By the definition (36) of α_{cvx} and the assumption $x_+ \in \mathcal{H}(\nabla f(x), \alpha_{\mathrm{cvx}}(x))$,

$$\langle \nabla f(x), x_+ \rangle \le \alpha_{\mathrm{cvx}}(x) = \langle \nabla f(x), x \rangle - f(x) + f_\star.$$

Hence, the Bregman divergence $D_f(x_+, x)$ can be bounded below as

$$\begin{aligned} D_f(x_+, x) &= f(x_+) - f(x) - \langle \nabla f(x), x_+ - x \rangle \\ &= f(x_+) - f_\star + (f_\star - f(x) + \langle \nabla f(x), x \rangle - \langle \nabla f(x), x_+ \rangle) \\ &\ge f(x_+) - f_\star. \end{aligned}$$

□

B.2 A universal rate for any method that cuts at the Polyak level

Lemma 16. *Suppose ∇f is Hölder continuous with exponent $\vartheta \in [0, 1]$ and constant L_ϑ, in the sense that*

$$D_f(u, v) \le \frac{L_\vartheta}{1+\vartheta}\|u - v\|^{1+\vartheta}, \qquad \forall u, v \in \mathbb{R}^d. \tag{69}$$

Fix $x_0 \in \mathcal{X}$ and let $\{x_k\}_{k \ge 0} \subset \mathcal{X}$ be generated by any method with the following two properties:

1. *$x_{k+1} \in \mathcal{H}(\nabla f(x_k), \alpha_{\mathrm{cvx}}(x_k))$ for all $k \ge 0$;*
2. *$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - \|x_{k+1} - x_k\|^2$ for all $k \ge 0$ and all $x_\star \in \mathcal{X}_\star$.*

Then, for every $K \ge 1$,

$$\min_{0 \le k \le K-1} (f(x_{k+1}) - f_\star) \le \frac{L_\vartheta}{1+\vartheta}\left(\frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}\right)^{\frac{1+\vartheta}{2}}. \tag{70}$$

Proof. By property 1, Lemma 15 applies at $x = x_k$ with $x_+ = x_{k+1}$. Combining (68) with the Hölder bound (69), used with $u = x_{k+1}$ and $v = x_k$, gives

$$f(x_{k+1}) - f_\star \le D_f(x_{k+1}, x_k) \le \frac{L_\vartheta}{1+\vartheta}\|x_{k+1} - x_k\|^{1+\vartheta}. \tag{71}$$

Property 2 telescopes as in the proof of Theorem 10 and yields

$$\min_{0 \le k \le K-1} \|x_k - x_{k+1}\|^2 \le \frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}, \qquad \forall K \ge 1. \tag{72}$$

Let $k^\star \in \mathrm{argmin}_{0 \le k \le K-1} \|x_{k+1} - x_k\|$. Applying (71) at $k = k^\star$ and then (72),

$$\begin{aligned} \min_{0 \le k \le K-1} (f(x_{k+1}) - f_\star) \;&\le\; f(x_{k^\star+1}) - f_\star \;\le\; \frac{L_\vartheta}{1+\vartheta}\|x_{k^\star+1} - x_{k^\star}\|^{1+\vartheta} \\ &\le\; \frac{L_\vartheta}{1+\vartheta}\left(\frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}\right)^{\frac{1+\vartheta}{2}}, \end{aligned}$$

which is (70). □

B.3 Proof of Theorem 3

Theorem 3 (Universal Hölder rate on the deep-cut band)**.** *Assume ∇f is Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ as in* (18)*. Let $\{x_k\}$ be generated from $x_0 \in \mathcal{X} \setminus \mathcal{X}_\star$ by $x_{k+1} = \mathrm{Proj}_{\mathcal{X} \cap \mathcal{H}_{x_k}(\nabla f(x_k),\, s_k)}(x_k)$ with any (possibly adaptive) depths in the deep-cut band, that is, satisfying $t(x_k) \le s_k \le s_\star(x_k)$ for all k, the run stopping if some $x_k \in \mathcal{X}_\star$. Then, for every $K \ge 1$, we have*

$$\min_{0 \le k \le K} \big(f(x_k) - f_\star\big) \le \frac{L_\vartheta}{1+\vartheta} \left(\frac{\mathrm{dist}^2(x_0, \mathcal{X}_\star)}{K}\right)^{\frac{1+\vartheta}{2}}. \tag{19}$$

Proof. Fix $K \ge 1$. If $x_k \in \mathcal{X}_\star$ for some $0 \le k \le K$, then $f(x_k) = f_\star$ and (19) holds trivially, so we may assume $x_k \in \mathcal{X} \setminus \mathcal{X}_\star$ for every $0 \le k \le K$.

We write $\alpha_k \coloneqq \langle g_k, x_k\rangle - s_k\|g_k\|$ for the level of the k-th cut, so that $\mathcal{H}_{x_k}(g_k, s_k) = \mathcal{H}(g_k, \alpha_k)$ by (32) and the hypothesis $t(x_k) \le s_k \le s_\star(x_k)$ reads $\alpha_\star(x_k) \le \alpha_k \le \alpha_{\mathrm{cvx}}(x_k)$ by (33). We verify the two hypotheses of Lemma 16. The lower bound $\alpha_k \ge \alpha_\star(x_k)$ gives $\mathcal{X}_\star \subseteq \mathcal{H}(g_k, \alpha_k)$ by Theorem 9, so Theorem 10 applies and the iterates satisfy property 2. The upper bound $\alpha_k \le \alpha_{\mathrm{cvx}}(x_k)$ gives $\mathcal{H}(g_k, \alpha_k) \subseteq \mathcal{H}(g_k, \alpha_{\mathrm{cvx}}(x_k))$, and since $x_{k+1} \in \mathcal{X} \cap \mathcal{H}(g_k, \alpha_k)$ by construction, property 1 holds as well. Lemma 16 then gives (70). Its minimum is taken over the K iterates $x_1, \dots, x_K$ produced by the method, so it is sharper than (19), whose minimum ranges over $x_0, \dots, x_K$ as well; we state the weaker form in the main text for consistency among (19), (21) and (95). □

C Proof of Theorem 4: a universal rate for Local LMO

Proof outline. Local LMO cuts above the Polyak level $\alpha_{\mathrm{cvx}}(x)$, so the analysis of Section B does not apply. For convenience, we define the suboptimality gap of iterate x_k as

$$a_k \coloneqq f(x_k) - f_\star. \tag{73}$$

We further define the projection of the initial iterate onto the set of minimizers

$$x_\star^0 \coloneqq \mathrm{Proj}_{\mathcal{X}_\star}(x_0), \tag{74}$$

and write $r_k \coloneqq \big\|x_k - x_\star^0\big\|$, so that $r_0 = R \coloneqq \mathrm{dist}(x_0, \mathcal{X}_\star)$. We also write $G_\star \coloneqq \|\nabla f(x_\star)\|$, which by Lemma 6 does not depend on the choice of $x_\star \in \mathcal{X}_\star$.

Our proof proceeds in four steps.

1. In Section C.1, we establish the conjugate Hölder inequality to give the gradient bound $\|g_k - \nabla f(x_\star)\| \le \kappa a_k^{\vartheta/(1+\vartheta)}$ of Lemma 19—whence, by the triangle inequality, the two-sided bound $\big|\|g_k\| - G_\star\big| \le \kappa a_k^{\vartheta/(1+\vartheta)}$ used in (96)—and, with Fejér monotonicity, a priori bounds that require no compactness.
2. Section C.2 tackles the local Frank–Wolfe gap and the overshoot. With $\ell_k \coloneqq \langle g_k, x_k - x_{k+1}\rangle$ and $\rho_k \coloneqq \ell_k / a_k \in (0,1]$, the descent inequality reads $a_{k+1} \le a_k(1 - \rho_k + \delta_k)$ with $\delta_k = c\, t_k^{1+\vartheta}/a_k$, and the overshoot $\Gamma_k = (1-\rho_k)a_k$ provides a floor on the next function value.
3. In Section C.3, we use the projection identity of Section A to tighten $r_{k+1}^2 \le r_k^2 - t_k^2$ to $r_{k+1}^2 \le r_k^2 - t_k^2(2-\rho_k)/\rho_k$, whence $\sum_{k<K} t_k^2/\rho_k \le R^2$. Here, $\rho_k \in (0,1]$ is the fraction of the available first-order decrease captured by the linear minimization oracle; at the Polyak radius $t_k = t(x_k)$ it is the cosine of the angle between the gradient $\nabla f(x_k)$ and the step $x_k - x_{k+1}$. The function value can fall at most a ρ_k fraction (Lemma 24), while the squared distance falls by t_k^2/ρ_k.
4. Section C.4 assembles the proof of Theorem 4. Let $A = \min_{k \le K} a_k$. If A is above the level $\bar a$ at which the Hölder part of the gradient dominates $G_\star$, the Fejér budget alone gives the exponent $\frac{1+\vartheta}{2}$. Below $\bar a$ the two-sided bound of step 1 pins $\|g_k\|$ to within a constant factor of $G_\star$, and the K steps are counted three ways (Cartis et al., 2011; 2012): those above $\bar a$ (few, by the budget); those that stall, $\rho_k < 2\delta_k$ (few, because each spends $a_k t_k^{1-\vartheta}/2c$ of the budget); and the productive rest, counted by Cauchy–Schwarz against the potential $\Phi_k = a_k^{-2}$, using the identity (91).

C.1 Two consequences of Hölder continuity

Throughout this section ∇f is Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ as in (18), and we write

$$c \coloneqq \frac{L_\vartheta}{1+\vartheta}, \quad \kappa_\vartheta \coloneqq \left(\frac{1+\vartheta}{\vartheta}\right)^{\frac{\vartheta}{1+\vartheta}}, \quad \kappa \coloneqq \kappa_\vartheta L_\vartheta^{\frac{1}{1+\vartheta}}, \quad h(a) \coloneqq \frac{a}{G_\star + \kappa\, a^{\vartheta/(1+\vartheta)}} \quad (a \geq 0), \tag{75}$$

with $h(0) \coloneqq 0$ when $G_\star = 0$. At the endpoint $\vartheta = 0$ we read κ_ϑ as its limit, $\kappa_0 \coloneqq \lim_{\vartheta \downarrow 0} \kappa_\vartheta = 1$. With this convention $\kappa = L_0$ and $h(a) = a/(G_\star + L_0)$ at $\vartheta = 0$, and all constants below are finite at both endpoints. Corollary 20 shows that h is increasing and that $t(x_k) \geq h(a_k)$ along any run.

The first lemma of this subsection is the Hölder analogue of the standard smooth inequality $D_f(u,v) \geq \frac{1}{2L}\|\nabla f(u) - \nabla f(v)\|^2$.

Lemma 17. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with $D_f(u,v) \leq \frac{L_\vartheta}{1+\vartheta}\|u-v\|^{1+\vartheta}$ for all u, v, where $\vartheta \in (0,1]$. Then*

$$D_f(u,v) \;\geq\; \frac{\vartheta}{1+\vartheta}\, L_\vartheta^{-1/\vartheta}\, \|\nabla f(u) - \nabla f(v)\|^{\frac{1+\vartheta}{\vartheta}} \qquad \forall u, v \in \mathbb{R}^d. \tag{76}$$

Proof. Fix v and set $\varphi(w) \coloneqq D_f(w,v) = f(w) - f(v) - \langle \nabla f(v), w - v\rangle$. Then φ is convex and differentiable with $\nabla\varphi(w) = \nabla f(w) - \nabla f(v)$ and $\varphi \geq 0$ by convexity of f. Since $\varphi = f - \ell$ with $\ell(w) \coloneqq f(v) + \langle \nabla f(v), w - v\rangle$ affine, and $D_\ell \equiv 0$ for affine ℓ because $\nabla\ell$ is constant, we have $D_\varphi = D_f - D_\ell = D_f$. Therefore, φ obeys the same Hölder bound, with the same L_ϑ.

Fix u and put $\Delta \coloneqq \|\nabla\varphi(u)\|$. If $\Delta = 0$, there is nothing to prove.

Otherwise, let, for $s > 0$, $w \coloneqq u - s\nabla\varphi(u)/\Delta$. Then

$$0 \leq \varphi(w) = \varphi(u) + \langle \nabla\varphi(u), w - u\rangle + D_\varphi(w,u) \leq \varphi(u) - s\Delta + \frac{L_\vartheta}{1+\vartheta} s^{1+\vartheta}.$$

Choosing $s = (\Delta/L_\vartheta)^{1/\vartheta}$ makes $\frac{L_\vartheta}{1+\vartheta} s^{1+\vartheta} = \frac{L_\vartheta s^\vartheta}{1+\vartheta}\, s = \frac{s\Delta}{1+\vartheta}$, so $0 \leq \varphi(u) - \frac{\vartheta}{1+\vartheta} s\Delta$, which is (76). □

The conjugate Hölder inequality is vacuous at $\vartheta = 0$. The following lemma supplies a Hölder-0 bound on the Bregman divergence which forces the gradient map to have bounded range.

Lemma 18. *Let f satisfy $D_f(u,v) \leq L_0\|u - v\|$ for all $u, v \in \mathbb{R}^d$. Then*

$$\|\nabla f(u) - \nabla f(v)\| \;\leq\; L_0, \qquad \forall u, v \in \mathbb{R}^d. \tag{77}$$

Proof. Fix v and set $\varphi(w) \coloneqq D_f(w,v)$, so that φ is convex and differentiable with $\nabla\varphi(w) = \nabla f(w) - \nabla f(v)$, $\varphi \geq 0$, and $D_\varphi = D_f$, exactly as in the proof of Lemma 17. Fix u and put $\Delta \coloneqq \|\nabla\varphi(u)\|$; if $\Delta = 0$ there is nothing to prove. For $s > 0$ let $w_s \coloneqq u - s\nabla\varphi(u)/\Delta$, so that $\|w_s - u\| = s$ and

$$0 \;\leq\; \varphi(w_s) \;=\; \varphi(u) + \langle \nabla\varphi(u), w_s - u\rangle + D_\varphi(w_s, u) \;\leq\; \varphi(u) - s\Delta + L_0\, s.$$

Hence $s\,(\Delta - L_0) \leq \varphi(u)$ for every $s > 0$. Were $\Delta > L_0$, the left-hand side would be unbounded above, a contradiction; so $\Delta \leq L_0$. □

Lemma 19. *Let ∇f be Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ. Then, for every $x \in \mathcal{X}$ and every $x_\star \in \mathcal{X}_\star$,*

$$\|\nabla f(x) - \nabla f(x_\star)\| \;\leq\; \kappa_\vartheta\, L_\vartheta^{\frac{1}{1+\vartheta}}\; (f(x) - f_\star)^{\frac{\vartheta}{1+\vartheta}}\,. \tag{78}$$

Proof. First-order optimality of $x_\star$ over $\mathcal{X}$ gives $\langle \nabla f(x_\star), x - x_\star\rangle \geq 0$ for all $x \in \mathcal{X}$, whence $D_f(x, x_\star) = f(x) - f_\star - \langle \nabla f(x_\star), x - x_\star\rangle \leq f(x) - f_\star$. Combining with Lemma 17 applied at $(u,v) = (x, x_\star)$, we find

$$\frac{\vartheta}{1+\vartheta} L_\vartheta^{-1/\vartheta} \|\nabla f(x) - \nabla f(x_\star)\|^{(1+\vartheta)/\vartheta} \leq f(x) - f_\star.$$

Raising to the power $\vartheta/(1+\vartheta)$ and rearranging gives (78). For $\vartheta = 0$, (78) reads $\|\nabla f(x) - \nabla f(x_\star)\| \leq \kappa_0 L_0 = L_0$, which is Lemma 18. □

Corollary 20. *Along any* Local LMO *run with* $0 < t_k \le t(x_k)$ *starting from* $x_0 \in \mathcal{X}$*, we have*

$$a_k \le a_{\max} := \max\left\{2G_\star R,\ (2\kappa_\vartheta)^{1+\vartheta} L_\vartheta R^{1+\vartheta}\right\} \tag{79}$$

$$\|g_k\| \le G := G_\star + \kappa_\vartheta L_\vartheta^{\frac{1}{1+\vartheta}} a_{\max}^{\frac{\vartheta}{1+\vartheta}} < +\infty, \tag{80}$$

and consequently

$$t(x_k) = \frac{a_k}{\|g_k\|} \ge \frac{a_k}{G_\star + \kappa\, a_k^{\vartheta/(1+\vartheta)}} = h(a_k), \tag{81}$$

with h *as in* (75)*, and* h *is strictly increasing on* $[0, +\infty)$*.*

Proof. Convexity and the Cauchy–Schwarz inequality give

$$a_k \le \left\langle g_k, x_k - x_\star^0 \right\rangle \le \|g_k\| r_k \le \|g_k\| R,$$

where the last step follows from Fejér monotonicity (part 3 of Lemma 14). With (78),

$$a_k \le G_\star R + \kappa_\vartheta L_\vartheta^{1/(1+\vartheta)} R\, a_k^{\vartheta/(1+\vartheta)}. \tag{82}$$

Now, if $a \le u + v a^p$ with $p = \frac{\vartheta}{1+\vartheta} \in [0, \frac{1}{2}]$ and $u, v \ge 0$, then either $a \le 2u$ or $a \le 2va^p$, i.e. $a \le (2v)^{1/(1-p)} = (2v)^{1+\vartheta}$. With $u = G_\star R$ and $v = \kappa R$ this is the first bound in (79), and the second follows from (78) again.

Equation (81) is just (78) together with $\|g_k\| \le G_\star + \|g_k - \nabla f(x_\star)\|$ for any $x_\star \in \mathcal{X}_\star$.

Finally $h(a) = a/(G_\star + \kappa a^p)$ with $p = \frac{\vartheta}{1+\vartheta} \in [0, \frac{1}{2}]$ has $h'(a) = \left(G_\star + (1-p)\kappa a^p\right)/(G_\star + \kappa a^p)^2 > 0$ for $a > 0$; if $G_\star = 0$ this reads $h(a) = a^{1/(1+\vartheta)}/\kappa$, which is continuous at 0 with our convention that $h(0) = 0$, so h is increasing on all of $[0, \infty)$ in either case. □

C.2 The local Frank–Wolfe gap and the overshoot

Definition 21. For $x \in \mathcal{X} \setminus \mathcal{X}_\star$ and $t > 0$, define

$$\ell(x, t) := \langle \nabla f(x), x\rangle - \min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle \nabla f(x), z\rangle = \langle \nabla f(x), x - x_{\mathrm{lmo}}\rangle \ \ge 0.$$

Lemma 22. *Fix* $x \in \mathcal{X} \setminus \mathcal{X}_\star$ *and* $0 < t \le t(x)$*. Then for every* $x_\star \in \mathcal{X}_\star$*,*

$$\ell(x, t) \ \ge\ \frac{t}{\|x - x_\star\|}\ \langle \nabla f(x), x - x_\star\rangle \ \ge\ \frac{t}{\|x - x_\star\|}\ (f(x) - f_\star)\,. \tag{83}$$

Proof. Let $g = \nabla f(x)$. As in steps 2 and 3 of the proof of Theorem 13, $0 < t \le t(x) \le \langle g, x - x_\star\rangle / \|g\| \le \|x - x_\star\|$, so $\lambda := t/\|x - x_\star\| \in (0, 1]$ and $z_t := x - \lambda(x - x_\star) = (1-\lambda)x + \lambda x_\star \in \mathcal{X}$ with $\|z_t - x\| = t$; hence $z_t \in \mathcal{X} \cap \mathcal{B}(x, t)$. Therefore

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, z\rangle \le \langle g, z_t\rangle = \langle g, x\rangle - \lambda\, \langle g, x - x_\star\rangle\,,$$

so $\ell(x, t) \ge \lambda\, \langle g, x - x_\star\rangle$. The last inequality in (83) is first-order convexity, $\langle g, x - x_\star\rangle \ge f(x) - f_\star$. □

Lemma 23. *Fix* $x_k \in \mathcal{X} \setminus \mathcal{X}_\star$ *and* $0 < t_k \le t(x_k)$*, and let* x_{k+1} *be a* Local LMO *step. With* $\ell_k := \ell(x_k, t_k)$*,*

$$\langle g_k, x_{k+1}\rangle = \alpha_{\mathrm{lmo}}(x_k, t_k) = \alpha_{\mathrm{cvx}}(x_k) + \Gamma_k, \qquad \Gamma_k := a_k - \ell_k = \alpha_{\mathrm{lmo}}(x_k, t_k) - \alpha_{\mathrm{cvx}}(x_k), \tag{84}$$

and Γ_k *satisfies*

$$0 \ \le\ \alpha_{\mathrm{proj}}(x_k, t_k) - \alpha_{\mathrm{cvx}}(x_k) \ \le\ \Gamma_k \ \le\ a_k \left(1 - \tfrac{t_k}{\|x_k - x_\star\|}\right), \qquad \forall x_\star \in \mathcal{X}_\star. \tag{85}$$

Consequently $x_{k+1} \in \mathcal{H}(g_k, \alpha_{\mathrm{proj}}(x_k, t_k))$ *if and only if* $\alpha_{\mathrm{lmo}}(x_k, t_k) = \alpha_{\mathrm{proj}}(x_k, t_k)$*, i.e. iff the unconstrained ball minimizer* $x_k - t_k\, g_k/\|g_k\|$ *is feasible in* $\mathcal{X}$*. Whenever the constraint* $\mathcal{X}$ *is active at the step,* $x_{k+1} \notin \mathcal{H}(g_k, \alpha_{\mathrm{proj}}(x_k, t_k))$*.*

Proof. By Theorem 13, $\langle g_k, x_{k+1}\rangle = \alpha_{\mathrm{lmo}}(x_k, t_k)$. By definition $\alpha_{\mathrm{cvx}}(x_k) = \langle g_k, x_k\rangle - a_k$, so $\alpha_{\mathrm{lmo}}(x_k, t_k) - \alpha_{\mathrm{cvx}}(x_k) = \langle g_k, x_{k+1}\rangle - \langle g_k, x_k\rangle + a_k = a_k - \ell_k = \Gamma_k$, which is (84). For (85): the middle inequality $\Gamma_k \geq \alpha_{\mathrm{proj}} - \alpha_{\mathrm{cvx}} \geq 0$ is Lemma 12 together with Lemma 11; the Cauchy–Schwarz bound $\ell_k = \langle g_k, x_k - x_{k+1}\rangle \leq \|g_k\|\,\|x_k - x_{k+1}\| = t_k\|g_k\| \leq t(x_k)\|g_k\| = a_k$ gives $\Gamma_k \geq 0$; and the upper bound is Lemma 22. The final claim is (84): $x_{k+1} \in \mathcal{H}(g_k, \alpha_{\mathrm{proj}}) \Leftrightarrow \alpha_{\mathrm{lmo}} \leq \alpha_{\mathrm{proj}} \Leftrightarrow \alpha_{\mathrm{lmo}} = \alpha_{\mathrm{proj}}$, using $\alpha_{\mathrm{lmo}} \geq \alpha_{\mathrm{proj}}$ (Lemma 12). Equality holds iff the minimum of $\langle g_k, \cdot\rangle$ over $\mathcal{B}(x_k, t_k)$, attained at $x_k - t_k g_k/\|g_k\|$, is attained inside $\mathcal{X}$. □

Lemma 24. *Under the assumptions of Lemma 23, every* Local LMO *step satisfies the exact identity* $f(x_{k+1}) - f_\star = D_f(x_{k+1}, x_k) + \Gamma_k$*, and hence, since* $D_f \geq 0$*,*

$$f(x_{k+1}) - f_\star \;\geq\; \Gamma_k \;=\; a_k - \ell_k \;\geq\; 0. \tag{86}$$

No Local LMO *step can push the suboptimality below the overshoot* Γ_k*, however smooth* f *is. By* (85)*,* $\Gamma_k \geq \alpha_{\mathrm{proj}}(x_k, t_k) - \alpha_{\mathrm{cvx}}(x_k) = \|g_k\|\,(t(x_k) - t_k)$*, so the floor vanishes only if* $t_k = t(x_k)$ and $\alpha_{\mathrm{lmo}} = \alpha_{\mathrm{proj}}$*, i.e. only if the radius is the full Polyak radius and the unconstrained ball minimizer is feasible — in which case the step coincides with the Polyak projection step. A shorter radius leaves a positive floor even when* $\mathcal{X}$ *is inactive.*

Proof. By the definition of the Bregman divergence and of ℓ_k,

$$D_f(x_{k+1}, x_k) = f(x_{k+1}) - f(x_k) - \langle g_k, x_{k+1} - x_k\rangle = f(x_{k+1}) - f(x_k) + \ell_k.$$

Adding and subtracting $f_\star$ and using $\Gamma_k = a_k - \ell_k$ (84) gives $f(x_{k+1}) - f_\star = D_f(x_{k+1}, x_k) + a_k - \ell_k = D_f(x_{k+1}, x_k) + \Gamma_k$. Since $D_f(x_{k+1}, x_k) \geq 0$ by convexity and $\Gamma_k \geq 0$ by (85), we obtain (86). □

Lemma 25. *Suppose* ∇f *is Hölder continuous with exponent* $\vartheta \in [0, 1]$ *(that is,* $D_f(u, v) \leq \frac{L_\vartheta}{1+\vartheta}\|u - v\|^{1+\vartheta}$ *for all* u, v*), and let* $\{x_k\}$ *be generated by* Local LMO *with* $0 < t_k \leq t(x_k)$*. With* $R = \mathrm{dist}(x_0, \mathcal{X}_\star)$*, for every* $k \geq 0$*,*

$$a_{k+1} \;\leq\; a_k - \ell_k + \frac{L_\vartheta}{1+\vartheta}\, t_k^{1+\vartheta} \;\leq\; \left(1 - \tfrac{t_k}{R}\right) a_k + \frac{L_\vartheta}{1+\vartheta}\, t_k^{1+\vartheta}. \tag{87}$$

Proof. The Hölder descent inequality with $u = x_{k+1}$, $v = x_k$ gives $a_{k+1} = a_k - \ell_k + D_f(x_{k+1}, x_k) \leq a_k - \ell_k + \frac{L_\vartheta}{1+\vartheta}\|x_{k+1} - x_k\|^{1+\vartheta}$, and $\|x_{k+1} - x_k\| = t_k$ by Theorem 13; this is the first inequality. For the second, take $x_\star = \mathrm{Proj}_{\mathcal{X}_\star}(x_0)$, so that part 3 of Lemma 14 gives $\|x_k - x_\star\| \leq \|x_0 - x_\star\| = R$, so Lemma 22 yields $\ell_k \geq (t_k/\|x_k - x_\star\|)a_k \geq (t_k/R)a_k$. □

C.3 The sharpened Fejér budget

Part 3 of Lemma 14 established the inequality $r_{k+1}^2 \leq r_k^2 - t_k^2$. Our next theorem shows that this is lossy when the overshoot Γ_k takes up a large share of the gap a_k, and quantifies the gain using the projection identity of Theorem 13.

Theorem 26. *Let* $x_k \in \mathcal{X} \setminus \mathcal{X}_\star$*, let* $0 < t_k \leq t(x_k)$*, and let* x_{k+1} *be a* Local LMO *step. Put* $\ell_k = \langle g_k, x_k - x_{k+1}\rangle$ (> 0) *and, for each* $x_\star \in \mathcal{X}_\star$*,*

$$\varepsilon_k(x_\star) \coloneqq \langle g_k, x_{k+1} - x_\star\rangle \geq \Gamma_k = a_k - \ell_k \geq 0.$$

Then, for every $x_\star \in \mathcal{X}_\star$*,*

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2\left(1 + \frac{2\varepsilon_k(x_\star)}{\ell_k}\right) \leq \|x_k - x_\star\|^2 - t_k^2\left(1 + \frac{2\Gamma_k}{\ell_k}\right). \tag{88}$$

Proof. Lemma 22 shows that $\ell_k \geq (t_k/r_k)a_k > 0$, with $r_k = \left\|x_k - x_\star^0\right\|$ as in (74).

To establish $\varepsilon_k(x_\star) \geq \Gamma_k \geq 0$, note that convexity gives $\langle g_k, x_k - x_\star\rangle \geq a_k$, i.e. $\langle g_k, x_\star\rangle \leq \alpha_{\mathrm{cvx}}(x_k)$. Hence, $\varepsilon_k(x_\star) = \alpha_{\mathrm{lmo}}(x_k, t_k) - \langle g_k, x_\star\rangle \geq \alpha_{\mathrm{lmo}}(x_k, t_k) - \alpha_{\mathrm{cvx}}(x_k) = \Gamma_k$ by Theorem 13 and (84), and $\Gamma_k \geq 0$ by Lemma 23.

Write $\alpha \coloneqq \alpha_{\mathrm{lmo}}(x_k, t_k)$ and $\mathcal{H}_k \coloneqq \mathcal{H}(g_k, \alpha)$, so that $\langle g_k, x_k\rangle = \alpha + \ell_k$ and $\langle g_k, x_\star\rangle = \alpha - \varepsilon_k(x_\star)$. Set

$$s \coloneqq \frac{\ell_k}{\ell_k + \varepsilon_k(x_\star)} \in (0,1], \qquad z \coloneqq (1-s)x_k + s x_\star.$$

Then $z \in \mathcal{X}$ by convexity, and

$$\langle g_k, z\rangle = (1-s)(\alpha+\ell_k) + s(\alpha - \varepsilon_k(x_\star)) = \alpha + \ell_k - s(\ell_k + \varepsilon_k(x_\star)) = \alpha,$$

so $z \in \mathcal{X} \cap \mathcal{H}_k$.

By Theorem 13, $x_{k+1} = \mathrm{Proj}_{\mathcal{X}\cap\mathcal{H}_k}(x_k)$, so the projection inequality at the point $z \in \mathcal{X}\cap\mathcal{H}_k$ reads $\langle x_k - x_{k+1}, z - x_{k+1}\rangle \le 0$. Since $z - x_{k+1} = (x_k - x_{k+1}) - s(x_k - x_\star)$ and $\|x_k - x_{k+1}\| = t_k$ (Theorem 13), this is

$$t_k^2 - s\,\langle x_k - x_{k+1}, x_k - x_\star\rangle \le 0, \qquad \text{i.e.} \qquad \langle x_k - x_{k+1}, x_k - x_\star\rangle \ge \frac{t_k^2}{s} = t_k^2\left(1 + \frac{\varepsilon_k(x_\star)}{\ell_k}\right).$$

Finally, $\|x_{k+1} - x_\star\|^2 = \|x_k - x_\star\|^2 - 2\,\langle x_k - x_{k+1}, x_k - x_\star\rangle + t_k^2$, which gives (88). The second inequality is $\varepsilon_k(x_\star) \ge \Gamma_k$. □

Remark 27. At the Polyak radius $t_k = t(x_k)$, with $x_k \in \mathcal{X}\setminus\mathcal{X}_\star$, one has $\ell_k \le \|g_k\| t_k = a_k$, so, specializing (88) to $x_\star = x_\star^0$ and writing $\rho_k \coloneqq \ell_k/a_k \in (0,1]$, (88) reads

$$r_{k+1}^2 \le r_k^2 - t_k^2\,\frac{2-\rho_k}{\rho_k} \le r_k^2 - \frac{t_k^2}{\rho_k}. \tag{89}$$

Lemma 24 now says that the step cannot reduce the value below $\Gamma_k = (1-\rho_k)a_k$: the smaller ρ_k, the worse the value progress. On the other hand, (89) says the smaller ρ_k, the better the distance progress, by precisely the reciprocal factor.

Corollary 28. *For* Local LMO *at the Polyak radius $t_k = t(x_k)$, with $x_k \in \mathcal{X}\setminus\mathcal{X}_\star$ for every $k < K$, for every $K \ge 1$,*

$$\sum_{k=0}^{K-1} \frac{t_k^2}{\rho_k} = \sum_{k=0}^{K-1} \frac{t_k^2\, a_k}{\ell_k} \le R^2. \tag{90}$$

Proof. This follows by telescoping (89) and using $r_K^2 \ge 0$. □

C.4 Proof of Theorem 4

The proof is a three-way accounting of the K steps, run against two budgets: the sharpened Fejér budget $\sum_{k<K} t_k^2/\rho_k \le R^2$ (Corollary 28) and the telescoping of the potential $\Phi_k \coloneqq a_k^{-2}$. What makes the two fit together is the following identity,

$$\underbrace{\frac{\rho_k}{a_k^2}}_{\text{potential gain}} \cdot \underbrace{\frac{t_k^2}{\rho_k}}_{\text{Fejér cost}} = \frac{t_k^2}{a_k^2} = \frac{1}{\|g_k\|^2}. \tag{91}$$

Gain times cost does not depend on ρ_k at all. A step that makes little progress (ρ_k small) pays for it in the Fejér budget by the reciprocal factor, so no choice of ρ_k is free, and Cauchy–Schwarz converts (91) into a bound on the number of steps. The only obstruction is the Hölder error $c\,t_k^{1+\vartheta}$, which can undo the potential gain. It is controlled by the gradient bound of Lemma 19, which together with the triangle inequality pins $\|g_k\|$ to within a factor of the constant $G_\star$ once a_k is below an explicit level, and by the Fejér budget above that level.

Throughout this section, we set $\lambda \coloneqq \frac{2}{3}$,

$$\bar a \coloneqq \left(\frac{\lambda G_\star}{\kappa}\right)^{\frac{1+\vartheta}{\vartheta}}, \qquad \text{so that} \qquad \kappa\,\bar a^{\,\frac{\vartheta}{1+\vartheta}} = \lambda G_\star, \tag{92}$$

with the conventions $\bar a \coloneqq 0$ if $G_\star = 0$ and $\bar a \coloneqq +\infty$ if $L_\vartheta = 0 < G_\star$, with $\kappa = \kappa_\vartheta L_\vartheta^{1/(1+\vartheta)}$ and $c = L_\vartheta/(1+\vartheta)$ as in (75), and

$$\delta_k \coloneqq \frac{c\,t_k^{1+\vartheta}}{a_k} = \frac{c\,t_k^{\vartheta}}{\|g_k\|}, \qquad \text{so that} \qquad a_{k+1} \le a_k\,(1 - \rho_k + \delta_k) \tag{93}$$

by Lemma 25 and $a_k = t_k\|g_k\|$.

Before we offer the proof of Theorem 4, we establish the bounds on the constants D_ϑ, E_ϑ introduced next.

With $\Theta_\vartheta := 5^{(1+\vartheta)/2}$, define

$$\begin{aligned} D_\vartheta &:= \left(\tfrac{5}{2}\,\kappa_\vartheta\right)^{1+\vartheta}, \qquad E_\vartheta := \max\left\{E_\vartheta^{(1)}, E_\vartheta^{(2)}, E_\vartheta^{(3)}\right\}, \\ E_\vartheta^{(1)} &:= \left[\tfrac{25}{9}\,2^\vartheta\left(\tfrac{3}{2}\right)^{1+\vartheta}\kappa_\vartheta^{1+\vartheta}\right]^{\frac{1}{2-\vartheta}}, \quad E_\vartheta^{(2)} := \left[\tfrac{8(1+\Theta_\vartheta)}{1+\vartheta}\left(\tfrac{5}{3}\right)^{1-\vartheta}\right]^{\frac{1}{2-\vartheta}}, \quad E_\vartheta^{(3)} := \tfrac{20}{3}. \end{aligned} \tag{94}$$

Lemma 29. *For every $\vartheta \in [0,1]$, $D_\vartheta \le \frac{25}{2}$ and $E_\vartheta \le 25$, both attained at $\vartheta = 1$.*

Proof. Write $\psi(\vartheta) := \kappa_\vartheta^{1+\vartheta} = \left(\frac{1+\vartheta}{\vartheta}\right)^\vartheta = \exp\left(\vartheta\log(1+\frac{1}{\vartheta})\right)$. Its exponent has derivative $\log(1+\frac{1}{\vartheta}) - \frac{1}{1+\vartheta} \ge 0$, because $\log(1+x) \ge \frac{x}{1+x}$ with $x = 1/\vartheta$ gives exactly $\log(1+\frac{1}{\vartheta}) \ge \frac{1/\vartheta}{1+1/\vartheta} = \frac{1}{1+\vartheta}$. So ψ is nondecreasing with $\psi(\vartheta) \le \psi(1) = 2$. Hence $D_\vartheta = (\frac{5}{2})^{1+\vartheta}\psi(\vartheta) \le (\frac{5}{2})^2 \cdot 2 = \frac{25}{2}$, attained at $\vartheta = 1$.

For E_ϑ note first that $\frac{1}{2-\vartheta} \le 1$, so a bracket that is ≥ 1 is not increased by raising it to that power; this settles the first bracket, while the second needs the sharper bound below. The first bracket is $\frac{25}{9}\,2^\vartheta(\frac{3}{2})^{1+\vartheta}\psi(\vartheta) \le \frac{25}{9}\cdot 2\cdot\frac{9}{4}\cdot 2 = 25$, with equality at $\vartheta = 1$, so $E_\vartheta^{(1)} \le 25$ and $E_1^{(1)} = 25$. For the second, put $B(\vartheta) := 8\left(1 + 5^{(1+\vartheta)/2}\right)(\frac{5}{3})^{1-\vartheta}/(1+\vartheta)$, so $E_\vartheta^{(2)} = B(\vartheta)^{1/(2-\vartheta)}$, and it suffices to show $B(\vartheta) \le 25^{\,2-\vartheta}$. Since $5^{(1+\vartheta)/2} \le 5$ on $[0,1]$ we have $B(\vartheta) \le \widetilde{B}(\vartheta) := 48(\frac{5}{3})^{1-\vartheta}/(1+\vartheta)$, and

$$\frac{d}{d\vartheta}\log\frac{25^{\,2-\vartheta}}{\widetilde{B}(\vartheta)} = \frac{1}{1+\vartheta} - \log 25 + \log\tfrac{5}{3} \;\le\; 1 - 3.218 + 0.512 \;<\; 0,$$

so $25^{2-\vartheta}/\widetilde{B}(\vartheta)$ is decreasing and attains its minimum on $[0,1]$ at $\vartheta = 1$, where it equals $25/48\cdot 2 = \frac{50}{48} > 1$. Hence $E_\vartheta^{(2)} \le \widetilde{B}(\vartheta)^{1/(2-\vartheta)} \le 25$. Finally, $E_\vartheta^{(3)} = \frac{20}{3} \le 25$. □

The endpoint $\vartheta = 0$ is settled separately, and more sharply, by the following elementary bound.

Proposition 30. *Let $\vartheta = 0$, that is, $D_f(u,v) \le L_0\|u - v\|$ for all $u, v \in \mathbb{R}^d$, and let $\{x_k\}$ be generated by* Local LMO *at the Polyak radius $t_k = t(x_k)$ from $x_0 \in \mathcal{X}$. Then, for every $K \ge 1$,*

$$\min_{0\le k\le K} a_k \;\le\; \frac{(G_\star + L_0)\,R}{\sqrt{K}}. \tag{95}$$

Proof. By Lemma 18 applied at $(u,v) = (x_k, x_\star)$ for any $x_\star \in \mathcal{X}_\star$, $\|g_k\| \le \|\nabla f(x_\star)\| + \|g_k - \nabla f(x_\star)\| \le G_\star + L_0$. At the Polyak radius $a_k = t_k\|g_k\|$, so $a_k \le (G_\star + L_0)\,t_k$ for every k. By part 3 of Lemma 14 and $r_0 = R$, telescoping gives $\sum_{k<K} t_k^2 \le r_0^2 - r_K^2 \le R^2$, hence $\min_{k<K} t_k^2 \le R^2/K$. Choosing $k^\star \in \operatorname{argmin}_{k<K} t_k$ (so that $k^\star \le K-1$, and the bound in fact holds for $\min_{0\le k\le K-1} a_k$),

$$\min_{0\le k\le K} a_k \;\le\; a_{k^\star} \;\le\; (G_\star + L_0)\,t_{k^\star} \;\le\; \frac{(G_\star + L_0)\,R}{\sqrt{K}}. \qquad \square$$

Theorem 4 (Universal Hölder rate for Local LMO)**.** *Let ∇f be Hölder continuous with exponent $\vartheta \in [0,1]$ and constant L_ϑ, and let $\{x_k\}$ be generated by* Local LMO *at the Polyak radius $t_k = t(x_k)$ from $x_0 \in \mathcal{X}\setminus\mathcal{X}_\star$, the run stopping if some $x_k \in \mathcal{X}_\star$. Then, for all $K \ge 1$, we have*

$$\min_{0\le k\le K}\big(f(x_k) - f_\star\big) \;\le\; \max\left\{\,\Lambda_1\,K^{-\frac{1+\vartheta}{2}}\;,\;\; \Lambda_2\,K^{-\frac{1}{2-\vartheta}}\,\right\}, \tag{21}$$

where $R = \operatorname{dist}(x_0, \mathcal{X}_\star)$ and $G_\star = \|\nabla f(x_\star)\|$ for any $x_\star \in \mathcal{X}_\star$, and

$$\Lambda_1 := \tfrac{25}{2}\,L_\vartheta R^{1+\vartheta}, \qquad \Lambda_2 := 25\max\Big\{\left(L_\vartheta R^2 G_\star^{1-\vartheta}\right)^{\frac{1}{2-\vartheta}},\; G_\star R\Big\}. \tag{22}$$

Proof. Suppose first that $\vartheta = 0$, so that both exponents in (21) equal $\frac{1}{2}$ and the claim reads $\min_{k\le K} a_k \le \max\{\Lambda_1, \Lambda_2\} K^{-1/2}$ with $\Lambda_1 = \frac{25}{2} L_0 R$ and $\Lambda_2 \ge 25\, G_\star R$, by (22). Since $u + v \le \max\left\{\frac{25}{2}u, 25v\right\}$ for all $u, v \ge 0$, Proposition 30 yields the claim. Assume from now on that $\vartheta \in (0, 1]$.

Set $A \coloneqq \min_{0\le k\le K} a_k$. If $A = 0$ there is nothing to prove, so assume $A > 0$; then $a_k \ge A > 0$ for every $k \le K$, and every t_k, ρ_k, δ_k below is well defined and finite. We use throughout: $\rho_k \in (0, 1]$ (Lemma 22 and Lemma 23); $t_k \ge h(a_k) \ge h(A)$ with h increasing (Corollary 20); $\sum_{k<K} t_k^2 \le \sum_{k<K} t_k^2/\rho_k \le R^2$ (Corollary 28); and the *two-sided* gradient bound

$$G_\star - \kappa\, a_k^{\frac{\vartheta}{1+\vartheta}} \;\le\; \|g_k\| \;\le\; G_\star + \kappa\, a_k^{\frac{\vartheta}{1+\vartheta}}, \tag{96}$$

which is (78) together with the triangle inequality in both directions.

Case 1: $A > \bar{a}$. We may assume $\kappa > 0$: if $L_\vartheta = 0$ then $\kappa = 0$, and either $G_\star > 0$, in which case $\bar{a} = +\infty$ by the convention following (92) and this case is empty, or $G_\star = 0$, in which case $D_f \equiv 0$ forces f to be affine with $\nabla f \equiv \nabla f(x_\star) = 0$, so $a_0 = 0$ and $A = 0$, already excluded.

By (92) this means $\kappa A^{\vartheta/(1+\vartheta)} > \lambda G_\star$, hence $G_\star + \kappa A^{\vartheta/(1+\vartheta)} < \frac{1+\lambda}{\lambda}\kappa A^{\vartheta/(1+\vartheta)}$ and

$$h(A) \;=\; \frac{A}{G_\star + \kappa A^{\vartheta/(1+\vartheta)}} \;>\; \frac{\lambda}{(1+\lambda)\kappa}\, A^{\frac{1}{1+\vartheta}}.$$

Since $t_k \ge h(A)$ for all $k < K$, the Fejér budget gives $K\, h(A)^2 \le R^2$, i.e. $A^{2/(1+\vartheta)} \le \left(\frac{(1+\lambda)\kappa}{\lambda}\right)^2 R^2/K$, i.e.

$$A \;\le\; \left(\tfrac{1+\lambda}{\lambda}\,\kappa R\right)^{1+\vartheta} K^{-\frac{1+\vartheta}{2}} \;=\; D_\vartheta\, L_\vartheta R^{1+\vartheta} K^{-\frac{1+\vartheta}{2}} \;\le\; \Lambda_1 K^{-\frac{1+\vartheta}{2}},$$

using $\kappa^{1+\vartheta} = \kappa_\vartheta^{1+\vartheta} L_\vartheta$ and $\frac{1+\lambda}{\lambda} = \frac{5}{2}$, so that the middle term is $\left(\frac{5}{2}\kappa_\vartheta\right)^{1+\vartheta} L_\vartheta R^{1+\vartheta} K^{-(1+\vartheta)/2} = D_\vartheta L_\vartheta R^{1+\vartheta} K^{-(1+\vartheta)/2}$ by (94), together with $D_\vartheta \le \frac{25}{2}$ from Lemma 29. (If $G_\star = 0$ then $\bar{a} = 0$ and this is the only case, so (21) holds with $\Lambda_2 = 0$.)

Case 2: $A \le \bar{a}$ (so $G_\star > 0$). Split $\{0, \dots, K-1\} = \mathcal{I} \cup \mathcal{J}$,

$$\mathcal{I} \coloneqq \{k < K :\, a_k \le \bar{a}\}, \qquad \mathcal{J} \coloneqq \{k < K :\, a_k > \bar{a}\}.$$

For $k \in \mathcal{I}$, (96) and $\kappa a_k^{\vartheta/(1+\vartheta)} \le \kappa \bar{a}^{\vartheta/(1+\vartheta)} = \lambda G_\star$ give the two-sided window

$$(1-\lambda)G_\star \;\le\; \|g_k\| \;\le\; (1+\lambda)G_\star, \qquad \text{hence} \qquad \frac{a_k}{(1+\lambda)G_\star} \;\le\; t_k \;\le\; \frac{a_k}{(1-\lambda)G_\star}. \tag{97}$$

For $k \in \mathcal{J}$ we have $t_k \ge h(a_k) > h(\bar{a}) = \bar{a}/\left((1+\lambda)G_\star\right)$, so $\sum_{k<K} t_k^2 \le R^2$ forces

$$|\mathcal{J}| \;\le\; K_1 \;\coloneqq\; \frac{(1+\lambda)^2 G_\star^2 R^2}{\bar{a}^{\,2}}. \tag{98}$$

Case 2a: $K \le 4K_1$. Since $A \le \bar{a}$ we have, as in (97), $h(A) \ge A/((1+\lambda)G_\star)$, so $K\, h(A)^2 \le R^2$ gives $A \le (1+\lambda)G_\star R\, K^{-1/2}$. As $\frac{1}{2-\vartheta} - \frac{1}{2} = \frac{\vartheta}{2(2-\vartheta)} \ge 0$ and $1 \le K \le 4K_1$,

$$A \;\le\; (1+\lambda)G_\star R\, K^{-\frac{1}{2-\vartheta}} K^{\frac{\vartheta}{2(2-\vartheta)}} \;\le\; (1+\lambda)G_\star R \left(\frac{2(1+\lambda)G_\star R}{\bar{a}}\right)^{\frac{\vartheta}{2-\vartheta}} K^{-\frac{1}{2-\vartheta}},$$

using $(4K_1)^{1/2} = 2(1+\lambda)G_\star R/\bar{a}$. By (92), $\bar{a}^{-\vartheta/(2-\vartheta)} = (\kappa/(\lambda G_\star))^{(1+\vartheta)/(2-\vartheta)}$, and since $1 + \frac{\vartheta}{2-\vartheta} = \frac{2}{2-\vartheta}$ the prefactor equals

$$(1+\lambda)^{\frac{2}{2-\vartheta}} 2^{\frac{\vartheta}{2-\vartheta}} \lambda^{-\frac{1+\vartheta}{2-\vartheta}} \kappa_\vartheta^{\frac{1+\vartheta}{2-\vartheta}} \left(G_\star^{1-\vartheta} R^2 L_\vartheta\right)^{\frac{1}{2-\vartheta}} \;=\; E_\vartheta^{(1)} \left(L_\vartheta R^2 G_\star^{1-\vartheta}\right)^{\frac{1}{2-\vartheta}} \;\le\; \Lambda_2,$$

because $(G_\star R)^{2/(2-\vartheta)} G_\star^{-(1+\vartheta)/(2-\vartheta)} = (G_\star^{1-\vartheta} R^2)^{1/(2-\vartheta)}$ and $\kappa^{(1+\vartheta)/(2-\vartheta)} = (\kappa_\vartheta^{1+\vartheta} L_\vartheta)^{1/(2-\vartheta)}$; with $\lambda = \frac{2}{3}$ the bracket is $(\frac{5}{3})^2 2^\vartheta (\frac{3}{2})^{1+\vartheta} \kappa_\vartheta^{1+\vartheta}$. The final inequality holds because $E_\vartheta^{(1)} \le E_\vartheta \le 25$ by Lemma 29, and $\Lambda_2 \ge 25\left(L_\vartheta R^2 G_\star^{1-\vartheta}\right)^{1/(2-\vartheta)}$ by (22).

Case 2b: $K > 4K_1$. Split $\mathcal{I} = \mathcal{I}_P \cup \mathcal{I}_S$ with

$$\mathcal{I}_P \coloneqq \{k \in \mathcal{I} :\ \rho_k \geq 2\delta_k\} \quad (\textit{productive}), \qquad \mathcal{I}_S \coloneqq \{k \in \mathcal{I} :\ \rho_k < 2\delta_k\} \quad (\textit{stalling}).$$

(i) Stalling steps are few. For $k \in \mathcal{I}_S$, $\rho_k < 2\delta_k$ and (93) give

$$\frac{t_k^2}{\rho_k} \ > \ \frac{t_k^2}{2\delta_k} \ = \ \frac{a_k\, t_k^{1-\vartheta}}{2c} \ \geq \ \frac{a_k^{2-\vartheta}}{2c\,((1+\lambda)G_\star)^{1-\vartheta}} \ \geq \ \frac{A^{2-\vartheta}}{2c\,((1+\lambda)G_\star)^{1-\vartheta}},$$

by the left inequality of (97) and $a_k \geq A$. Summing and using Corollary 28,

$$|\mathcal{I}_S| \ \leq \ N \ \coloneqq \ 2c\,((1+\lambda)G_\star)^{1-\vartheta}\, R^2 A^{\vartheta-2}. \tag{99}$$

(ii) The potential. Put $\Phi_k \coloneqq a_k^{-2}$, well defined since $a_k \geq A > 0$. If $k \in \mathcal{I}_P$ then $\rho_k \geq 2\delta_k$ and (93) give $a_{k+1} \leq a_k(1 - \frac{\rho_k}{2})$ with $1 - \frac{\rho_k}{2} \geq \frac{1}{2} > 0$, so, by $(1-s)^{-2} \geq 1 + 2s$ on $[0, 1)$ applied at $s = \rho_k/2$,

$$\Phi_{k+1} - \Phi_k \ \geq \ \Phi_k \left[(1 - \tfrac{\rho_k}{2})^{-2} - 1\right] \ \geq \ \frac{\rho_k}{a_k^2}. \tag{100}$$

If $k \in \mathcal{I}_S$ then $a_{k+1} \leq a_k(1 + \delta_k)$, so by $1 - (1+s)^{-2} \leq 2s$ on $[0, \infty)$ and, by the right inequality of (97), $\delta_k = c\, t_k^{\vartheta}/\|g_k\| \leq c\, a_k^{\vartheta}/\left((1-\lambda)^{1+\vartheta} G_\star^{1+\vartheta}\right)$,

$$\Phi_k - \Phi_{k+1} \ \leq \ \frac{2\delta_k}{a_k^2} \ \leq \ \frac{2c\, a_k^{\vartheta-2}}{(1-\lambda)^{1+\vartheta} G_\star^{1+\vartheta}} \ \leq \ \frac{2c\, A^{\vartheta-2}}{(1-\lambda)^{1+\vartheta} G_\star^{1+\vartheta}}, \tag{101}$$

the last step because $\vartheta - 2 < 0$ and $a_k \geq A$. If $k \in \mathcal{J}$ then simply $\Phi_k - \Phi_{k+1} \leq \Phi_k = a_k^{-2} < \bar{a}^{-2}$. Telescoping over all $k < K$ and discarding $-\Phi_0 \leq 0$,

$$\sum_{k \in \mathcal{I}_P} (\Phi_{k+1} - \Phi_k) \ = \ \Phi_K - \Phi_0 - \sum_{k \in \mathcal{I}_S \cup \mathcal{J}} (\Phi_{k+1} - \Phi_k) \ \leq \ \Phi_K + \sum_{k \in \mathcal{I}_S \cup \mathcal{J}} (\Phi_k - \Phi_{k+1}),$$

so, with $\Phi_K \leq A^{-2}$, (100), (101) and (98),

$$\sum_{k \in \mathcal{I}_P} \frac{\rho_k}{a_k^2} \ \leq \ \frac{1}{A^2} \ + \ \frac{2c\,|\mathcal{I}_S| A^{\vartheta-2}}{(1-\lambda)^{1+\vartheta} G_\star^{1+\vartheta}} \ + \ \frac{K_1}{\bar{a}^{\,2}}. \tag{102}$$

(iii) Productive steps are few. By Cauchy–Schwarz, Corollary 28 and $t_k/a_k = 1/\|g_k\| \geq 1/\left((1+\lambda)G_\star\right)$ on $\mathcal{I}$,

$$\left(\frac{|\mathcal{I}_P|}{(1+\lambda)G_\star}\right)^2 \leq \left(\sum_{k \in \mathcal{I}_P} \frac{t_k}{a_k}\right)^2 \leq \left(\sum_{k \in \mathcal{I}_P} \frac{t_k^2}{\rho_k}\right)\left(\sum_{k \in \mathcal{I}_P} \frac{\rho_k}{a_k^2}\right) \leq R^2 \sum_{k \in \mathcal{I}_P} \frac{\rho_k}{a_k^2}.$$

Insert (102) and use $\sqrt{x+y+z} \leq \sqrt{x} + \sqrt{y} + \sqrt{z}$. The first term contributes $(1+\lambda)G_\star R/A$; the third contributes $(1+\lambda)G_\star R\sqrt{K_1}/\bar{a} = \left((1+\lambda)G_\star R/\bar{a}\right)^2 = K_1$; and the second, using $|\mathcal{I}_S| \leq N$ from (99), contributes

$$(1+\lambda)G_\star R \cdot \frac{\sqrt{2cN A^{\vartheta-2}}}{(1-\lambda)^{\frac{1+\vartheta}{2}} G_\star^{\frac{1+\vartheta}{2}}} = 2cR^2 G_\star^{1-\vartheta} A^{\vartheta-2} \frac{(1+\lambda)^{\frac{3-\vartheta}{2}}}{(1-\lambda)^{\frac{1+\vartheta}{2}}} = \Theta\, N, \qquad \Theta \coloneqq \left(\frac{1+\lambda}{1-\lambda}\right)^{\frac{1+\vartheta}{2}},$$

by (99) again. Altogether $|\mathcal{I}_P| \leq (1+\lambda)G_\star R/A + \Theta N + K_1$, and therefore

$$K \ = \ |\mathcal{J}| + |\mathcal{I}_S| + |\mathcal{I}_P| \ \leq \ 2K_1 + (1+\Theta)N + \frac{(1+\lambda)G_\star R}{A}. \tag{103}$$

(iv) Conclusion. Since $K > 4K_1$ we have $2K_1 < K/2$, so (103) gives $(1+\Theta)N + (1+\lambda)G_\star R/A > K/2$ and at least one of the two terms is $\geq K/4$.

If $(1+\Theta)N \geq K/4$ then, by (99) and $c = L_\vartheta/(1+\vartheta)$,

$$A^{2-\vartheta} \ \leq \ \frac{8(1+\Theta)(1+\lambda)^{1-\vartheta}}{1+\vartheta} \cdot \frac{L_\vartheta R^2 G_\star^{1-\vartheta}}{K}, \qquad \text{i.e.} \qquad A \ \leq \ E_\vartheta^{(2)} \left(L_\vartheta R^2 G_\star^{1-\vartheta}\right)^{\frac{1}{2-\vartheta}} K^{-\frac{1}{2-\vartheta}},$$

which is at most $\Lambda_2 K^{-1/(2-\vartheta)}$, because $E_\vartheta^{(2)} \le E_\vartheta \le 25$ by Lemma 29; with $\lambda = \frac{2}{3}$ one has $\Theta = 5^{(1+\vartheta)/2}$ and $1+\lambda = \frac{5}{3}$.

If instead $(1+\lambda)G_\star R/A \ge K/4$ then $A \le 4(1+\lambda)G_\star R\, K^{-1} \le E_\vartheta^{(3)}\, G_\star R\, K^{-1/(2-\vartheta)}$, since $1 \ge \frac{1}{2-\vartheta}$ and $K \ge 1$; this too is at most $\Lambda_2 K^{-1/(2-\vartheta)}$, since $E_\vartheta^{(3)} = \frac{20}{3} \le 25$ (Lemma 29) and $\Lambda_2 \ge 25\, G_\star R$ by (22).

In every case $A \le \max\{\Lambda_1 K^{-(1+\vartheta)/2}, \Lambda_2 K^{-1/(2-\vartheta)}\}$, which is (21). □

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks:

1. **Aid or polish writing.** An LLM has been employed for proof-reading and suggesting edits to enhance clarity of exposition, as well as formatting. The final wording reflects our own preferences for the presentation of our work.
2. **Retrieval and discovery.** An LLM has been employed to identify additional related work. All cited works have been reviewed by the authors for relevance.
3. **Research ideation or execution.** An LLM has been employed during ideation. The discovery of Theorem 4 and its proof have been significantly supported by AI following the human-led discovery of Theorems 2 and 3.
4. **Generating synthetic datasets and coding.** An LLM has been employed to come up with problem instances to test our theory on, including the one presented in Section 4. LLM-generated code was verified and tested for correctness by the authors.
5. **Proving mathematical claims.** An LLM has been employed to draw up the proof of Theorem 4 at the authors' direction. All proofs have been compiled by the authors and reviewed for correctness and clarity of exposition.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.