A FEW ACCELERATED ALGORITHMS FOR CONVEX OPTIMIZATION UNDER (H_0, H_1)-SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop accelerated algorithms for convex (H_0, H_1)-smooth optimization, where $\left\|\nabla^2 f(x)\right\| \leq H_0 + H_1(f(x) - f^*)$. This class generalizes standard smoothness and contains the (L_0, L_1)-smooth class. Combining a Nesterov-type accelerated gradient scheme with small-dimensional relaxation and phase restarts, we obtain a full-gradient method with iteration complexity $\widetilde{O}(\sqrt{H_0 \widetilde{R}^2/\varepsilon} + \sqrt{H_1 \widetilde{R}^2}\log(F_0/\varepsilon))$. We extend the same approach to randomized coordinate optimization, obtaining a coordinate method with uniform sampling whose iteration complexity carries the standard factor d, and a coordinate method with non-uniform sampling whose iteration complexity is governed by $S_{1/2}^{(j)} = \sum_i \sqrt{H_{j,i}}$. These results provide, to our knowledge, the first accelerated full-gradient and coordinate guarantees for this convex class. We also provide practical implementation recommendations. Experiments confirm the predicted acceleration, gains from non-uniform sampling, and the viability of inexact relaxation.

1 INTRODUCTION

Gradient descent is one of the basic computational primitives of modern optimization. Its stochastic and adaptive descendants, including SGD, momentum, AdaGrad, Adam, AdamW, and more recent matrix-valued optimizers such as Muon, are the default tools behind large-scale machine learning (Bottou et al., 2018; Duchi et al., 2011; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Liu et al., 2025a). Even when these methods differ in normalization, adaptivity, momentum, or matrix preconditioning, they share the same first-order nature: each iteration uses local descent information and turns it into a parameter update. This is why a sharper understanding of gradient-based optimization assumptions often translates directly into sharper theory for practical training algorithms.

The importance of this question is particularly visible in deep learning and large language model training. Modern architectures are trained almost entirely with first-order methods and their adaptive variants, from Transformers (Vaswani et al., 2017) to large-scale language models (Brown et al., 2020). At the same time, successful training recipes rely on mechanisms that are not naturally explained by the classical globally smooth model: gradient clipping is used to stabilize training with large or transient gradients (Pascanu et al., 2013; Zhang et al., 2020b), learning-rate warmup is standard in large-batch and Transformer training (Goyal et al., 2017; Vaswani et al., 2017), and clipped or normalized updates are particularly effective in overparameterized regimes (Lobanov & Koloskova, 2026). These empirical practices suggest that relevant curvature is not a single global constant. It changes along the trajectory and often becomes much smaller as training approaches a good solution.

Classical accelerated theory gives optimal guarantees for convex optimization under a global Lipschitz bound on the gradient (Nesterov, 1983; 2018). This assumption is sharp as a worst-case abstraction, but it can be too conservative for objectives whose curvature decreases during training. This gap between practice and theory motivated generalized smoothness assumptions. The most studied example is (L_0, L_1)-smoothness, $\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$, which was introduced to model the observed dependence of local smoothness on the gradient norm in deep learning and was later developed as a framework for clipping, normalization, and convex optimization under generalized smoothness (Zhang et al., 2020b; Koloskova et al., 2023; Li et al., 2023a; Lobanov et al., 2024; Vankov et al., 2025; Tyurin, 2024; Gorbunov et al., 2025; Lobanov & Gasnikov, 2025;

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.

Gaash et al., 2025; Tyurin, 2025; Borodich & Kovalev, 2026; Lobanov & Koloskova, 2026). More recently, gap-dependent smoothness assumptions, in particular (H_0, H_1)-smoothness assumptions, were proposed, where the curvature is controlled by the current optimality gap rather than by the gradient norm. Following this line of work (Vaswani & Babanezhad, 2025; Liu et al., 2025b; Alimisis et al., 2025; Lobanov & Koloskova, 2026; Vaswani et al., 2026), in this paper we focus on convex objectives satisfying the (H_0, H_1)-smoothness condition

$$\left\|\nabla^2 f(x)\right\| \le H_0 + H_1\big(f(x) - f^*\big), \qquad H_0, H_1 \ge 0. \tag{1}$$

When $H_1 = 0$, this is the standard H_0-smooth convex setting, while $H_1 > 0$ allows the curvature to be large far from the optimum and to decrease with the optimization gap. Importantly, this model is not disjoint from the (L_0, L_1)-smooth literature: it covers the (L_0, L_1)-smooth class and is strictly broader, since there are (H_0, H_1)-smooth functions that are not (L_0, L_1)-smooth; see Section B, Propositions B.1, B.4, and B.5 of Alimisis et al. (2025).

The strict inclusion makes acceleration under (H_0, H_1)-smoothness distinct from the existing (L_0, L_1)-smooth theory. For the latter class, Vankov et al. (2025) obtain $\widetilde{O}(\sqrt{L_0 R^2/\varepsilon} + (L_1 R)^{2/3}\log(F_0/\varepsilon))$, while the guarantee of Gorbunov et al. (2025) involves an exponential dependence on $L_1 R$. Neither result yields acceleration in terms of H_0 and H_1, leading to our central question:

Can one obtain accelerated rates for the broader convex (H_0, H_1)-smooth class, and how should the complexity depend on H_0 and H_1?

A naive answer would be to freeze the curvature on the initial sublevel set. With $\widetilde{R} := \sup_{x: f(x) \le f(x^0)} \|x - x^*\|$ and $L_{\mathrm{lev}} := H_0 + H_1 F_0$, classical acceleration gives

$$O\left(\widetilde{R}\sqrt{\frac{H_0 + H_1 F_0}{\varepsilon}}\right).$$

This bound ignores the fact that the $H_1 F(x)$ part should decrease as the method approaches the solution set. Our goal is to separate the baseline curvature H_0 from the gap-dependent curvature $H_1 F(x)$ and obtain, up to logarithmic and absolute factors,

$$\widetilde{O}\left(\widetilde{R}\left(\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\log\frac{F_0}{\varepsilon}\right)\right).$$

This bound has the classical accelerated dependence on the baseline curvature H_0. The H_1-term also has an accelerated, strongly-convex-like form: it contributes $\widetilde{R}\sqrt{H_1}\log(F_0/\varepsilon)$, namely a logarithmic dependence on the desired accuracy, up to the level-set radius factor $\widetilde{R}$. In particular, the rate avoids the $\sqrt{F_0/\varepsilon}$ penalty that would arise from freezing the curvature at the initial level set.

We also study the coordinate version of this question. Coordinate descent methods are designed for large-scale problems in which a full gradient is too expensive, while a single partial derivative or a small block update can be computed cheaply and can exploit sparsity or separable structure. This granularity is central in huge-scale convex optimization and has led to a rich accelerated theory under global coordinate smoothness (Nesterov, 2012; Richtarik & Takac, 2014; Fercoq & Richtarik, 2015; Allen-Zhu et al., 2016). However, these guarantees rely on fixed coordinate Lipschitz constants. To the best of our knowledge, accelerated results for convex (L_0, L_1)-smooth optimization were not previously available for coordinate methods. Since the (H_0, H_1)-smooth model strictly contains the (L_0, L_1)-smooth class, the coordinate case poses the sharper challenge of obtaining acceleration in the broader gap-dependent smoothness regime, namely under (H_0, H_1)-smoothness:

Can one obtain coordinate acceleration without replacing the problem by pessimistic coordinate smoothness constants on the initial level set?

In answering this question, we aim for the coordinate analogue of the full-gradient accelerated rate: the same (H_0, H_1)-dependence, up to the standard factor d for using one partial derivative per iteration. Thus a randomized coordinate method should use

$$\widetilde{O}\left(d\widetilde{R}\left(\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\log\frac{F_0}{\varepsilon}\right)\right)$$

coordinate-gradient calls, preserving acceleration in H_0 and the strongly-convex-like logarithmic H_1-term.

Table 1: Comparison of the iteration complexity ($\#N$) of state-of-the-art full-gradient and coordinate algorithms. Notation: d is the problem dimension; $S_{1/2} = \sum_{i=1}^{d} \sqrt{L_i}$ (L_i are Lipschitz constants for partial derivatives); $\widetilde{R} = \sup_{x \in \mathbb{R}^d : f(x) \le f(x_0)} \|x - x^*\|$; ε is the desired solution accuracy; $R = \|x_0 - x^*\|$; $F_0 = f(x_0) - f(x^*)$; FM stands for a full-gradient method; CM stands for a coordinate method.

Reference	$\#N$	Method		(L_0, L_1)?	(H_0, H_1)?	Acceleration?
		FM?	CM?			
(Nesterov & Stich, 2017)	$\mathcal{O}\left(\sqrt{\frac{S_{1/2}^2 \widetilde{R}^2}{\varepsilon}}\right)$	✗	✓	✗	✗	✓
(Lobanov et al., 2024)	$\mathcal{O}\left(\max\left\{\frac{dL_0R^2}{\varepsilon}, dL_1R\log\frac{F_0}{\varepsilon}\right\}\right)$	✗	✓	✓	✗	✗
(Lobanov et al., 2024) & Vankov et al. (2025)	$\mathcal{O}\left(\frac{L_0R^2}{\varepsilon} + L_1R\log\frac{F_0}{\varepsilon}\right)$	✓	✗	✓	✗	✗
Vankov et al. (2025)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L_0R^2}{\varepsilon}} + (L_1R)^{2/3}\log\frac{F_0}{\varepsilon}\right)$	✓	✗	✓	✗	✓
(Lobanov & Koloskova, 2026)	$\mathcal{O}\left(\frac{H_0R^2}{\varepsilon} + H_1R^2\log\frac{F_0}{\varepsilon}\right)$	✓	✗	✓	✓	✗
This paper	$\tilde{\mathcal{O}}\left(\sqrt{\frac{H_0\widetilde{R}^2}{\varepsilon}} + \sqrt{H_1\widetilde{R}^2}\log\frac{F_0}{\varepsilon}\right)$	✓	✗	✓	✓	✓
	$\tilde{\mathcal{O}}\left(d\sqrt{\frac{H_0\widetilde{R}^2}{\varepsilon}} + d\sqrt{H_1\widetilde{R}^2}\log\frac{F_0}{\varepsilon}\right)$	✗	✓	✓	✓	✓

2 Related Work and Main Contributions

The literature on gradient-descent variants under generalized smoothness is already extensive and remains very active. We discuss below the contributions most relevant to our work and display a representative selection in Table 1.

Full-gradient methods. Although our results are for convex objectives, the generalized smoothness line was first motivated by non-convex training. In particular, Zhang et al. (2020b) introduced (L_0, L_1)-smoothness for twice differentiable objectives, $\|\nabla^2 f(x)\| \le L_0 + L_1 \|\nabla f(x)\|$, as a relaxation of global L-smoothness motivated by the empirical correlation between local curvature and gradient norm during neural-network training. This condition was later reformulated in a first-order form by Zhang et al. (2020a), extending the (L_0, L_1)-smooth framework to differentiable functions that need not be twice differentiable; subsequent work developed this direction in non-convex and stochastic settings (Crawshaw et al., 2022; Chen et al., 2023; Wang et al., 2023; Koloskova et al., 2023; Li et al., 2023a;b; Hübler et al., 2024; Vankov et al., 2025). For convex (L_0, L_1)-smooth objectives, the non-accelerated theory has progressed from results that still relied on an additional global L-smoothness assumption (Koloskova et al., 2023; Takezawa et al., 2024), to bounds without global smoothness (Gorbunov et al., 2025; Lobanov et al., 2024; Vankov et al., 2025). Gorbunov et al. (2025) obtained $\mathcal{O}\left(L_0R^2/\varepsilon + L_1^2R^2\right)$, where the second term does not depend on the desired accuracy. This was sharpened by Lobanov et al. (2024); Vankov et al. (2025) to the mixed rate $\mathcal{O}\left(L_0R^2/\varepsilon + L_1R\log(F_0/\varepsilon)\right)$. Accelerated methods under (L_0, L_1)-smoothness were studied by Li et al. (2023a); Gorbunov et al. (2025); Vankov et al. (2025). The first two bounds couple L_0 and L_1 or involve exponential or level-set factors, whereas Vankov et al. (2025) separates the two contributions and obtains $\widetilde{O}(\sqrt{L_0R^2/\varepsilon} + (L_1R)^{2/3}\log(F_0/\varepsilon))$. These results characterize acceleration in the native (L_0, L_1) parametrization and do not directly provide a rate in terms of H_0 and H_1. The gap-dependent (H_0, H_1)-smoothness model, $\|\nabla^2 f(x)\| \le H_0 + H_1(f(x) - f^*)$, was recently studied in (Vaswani & Babanezhad, 2025; Liu et al., 2025b; Alimisis et al., 2025) and is weaker than the (L_0, L_1)-smooth model. Existing convex guarantees in this regime are non-accelerated: Lobanov & Koloskova (2026) obtains $\mathcal{O}\left(H_0R^2/\varepsilon + H_1R^2\log(F_0/\varepsilon)\right)$. To the best of our knowledge, no accelerated full-gradient rate was previously known for convex (H_0, H_1)-smooth objectives. Our result gives the first such guarantee, $\tilde{\mathcal{O}}\left(\sqrt{H_0\widetilde{R}^2/\varepsilon} + \sqrt{H_1\widetilde{R}^2}\log(F_0/\varepsilon)\right)$, separating H_0 and H_1 and requiring no additional global smoothness.

Coordinate methods. Coordinate descent methods are a natural alternative to full-gradient methods in large-scale problems, where reading all coordinates of the gradient at every iteration may be substantially more expensive than querying a single partial derivative. Their classical theory is well developed under fixed coordinate smoothness constants L_i: randomized coordinate descent and its accelerated, parallel, proximal, and non-uniformly sampled variants achieve rates governed by coordinate smoothness summaries such as $S_{1/2} = \sum_i \sqrt{L_i}$ (Nesterov, 2012; Richtarik & Takac, 2014; Fercoq & Richtarik, 2015; Nesterov & Stich, 2017; Allen-Zhu et al., 2016). These guarantees are the

coordinate analogue of classical smooth optimization: they are sharp in the global model, but they require global coordinate Lipschitz constants and therefore inherit the same pessimism when local curvature is better described by generalized smoothness. Much less is known in this regime. For convex (L_0, L_1)-smooth objectives, Lobanov et al. (2024) obtained a coordinate method with complexity $\mathcal{O}\left(\max\{dL_0R^2/\varepsilon, dL_1R\log(F_0/\varepsilon)\}\right)$, but this is a non-accelerated rate. To the best of our knowledge, no previous coordinate method is both accelerated and applicable to convex (L_0, L_1)-smooth objectives, and no accelerated coordinate result was known for the broader (H_0, H_1)-smooth class. Our coordinate result fills this gap: it gives an accelerated randomized coordinate method under convex (H_0, H_1)-smoothness, applies in particular to the (L_0, L_1)-smooth subclass, and pays only the natural additional factor d in coordinate-gradient calls relative to the full-gradient rate.

3 Problem setup

We consider the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{2}$$

where the objective $f : \mathbb{R}^d \to \mathbb{R}$ maps a parameter vector to a scalar loss value. Throughout the paper we use the Euclidean norm for vectors and the induced spectral norm for matrices.

We assume that the solution set $X^* := \arg\min_{x\in\mathbb{R}^d} f(x)$ is nonempty and bounded. We fix an arbitrary minimizer $x^* \in X^*$ and write $f^* := f(x^*)$, $F(x) := f(x) - f^*$, and $F_0 := F(x^0)$, where x^0 is the initial point.

Assumption 1 (Convexity)**.** *The function f is differentiable and convex, that is,*

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

As discussed above, this paper focuses on (H_0, H_1)-smoothness (Liu et al., 2025b; Alimisis et al., 2025; Lobanov & Koloskova, 2026).

Assumption 2 ((H_0, H_1)-smoothness)**.** *There exist constants $H_0 \geq 0$ and $H_1 > 0$. Let $\mathcal{L}(\Delta) := 2(H_0 + H_1\Delta)$ and $r_1 := 1/((2\sqrt{3} + \sqrt{6})\sqrt{H_1})$. For all $x, y \in \mathbb{R}^d$ satisfying $\|y - x\| \leq r_1$,*

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

In Assumption 2, the limiting case $H_1 = 0$ is the usual globally smooth convex setting: one can take $\mathcal{L}(\Delta) = 2H_0$ and $r_1 = \infty$. The pure H_1-driven regime $H_0 = 0$ is covered when a minimizer exists. For positive objectives where a solution may fail to exist, one should instead use an uncentered positive-function variant, with the curvature controlled by $f(x)$ rather than $f(x) - f^*$, as in Vaswani & Babanezhad (2025); Vaswani et al. (2026).

Assumption 2 is a first-order local-model formulation and does not require second derivatives. For twice continuously differentiable convex functions, it corresponds, up to absolute constants, to $\left\|\nabla^2 f(x)\right\| \leq H_0 + H_1(f(x) - f^*)$; see Liu et al. (2025b, Lemma 2 and Appendix C.2).

Moreover, Assumption 2 is strictly weaker than the (L_0, L_1)-smooth framework. First, Proposition B.1 (with $v = 1$) of Alimisis et al. (2025) shows that every (L_0, L_1)-smooth function satisfies Assumption 2 with $H_0 = L_0 + L_0L_1$ and $H_1 = (4L_1^2 + L_1)/2$. Second, Examples 3 and 4 are convex (H_0, H_1)-smooth functions but fail to satisfy any (L_0, L_1)-smoothness condition with finite constants. The proofs are deferred to Appendix B.

Example 3 (One-dimensional)**.** *Fix $p \in (1, 2]$. Let $q_p(t) := 1+(1+t^2)^{p/2}\exp(-(1+t^2)^p\sin^2(\pi t))$, and define*

$$\phi_p(x) = \int_0^x \int_0^s q_p(t)\, dt\, ds.$$

Then ϕ_p is convex and satisfies Assumption 2 with $H_0 = H_1 = 2$, but is not (L_0, L_1)-smooth for any finite L_0, L_1.

Example 4 (Multidimensional)**.** *For $d \geq 2$, let*

$$\Phi_p(x_1, \ldots, x_d) = \phi_p(x_1) + \frac{1}{2}\sum_{i=2}^{d} x_i^2,$$

where ϕ_p is the function from Example 3. Then Φ_p is strongly convex and satisfies Assumption 2 with $H_0 = H_1 = 2$, but is not (L_0, L_1)-smooth for any finite L_0, L_1.

Examples 3 and 4 contain narrow high-curvature peaks. The Hessian can be large, while the gradient, which integrates curvature over a neighborhood, remains much smaller. This separates (H_0, H_1)-smoothness from assumptions that force curvature to scale with $\|\nabla f(x)\|$. In machine learning tasks, where losses may be highly anisotropic and locally sensitive along only a few parameter directions, this distinction is useful because local sensitivity and accumulated first-order signal need not vary at the same rate.

4 Accelerated Algorithms

We now begin the presentation of our main algorithmic results. We first introduce a common outer wrapper (see Algorithm 1) that will be instantiated with both the full-gradient and coordinate inner methods developed below.

Algorithm 1 Restart Meta-Algorithm

Input: objective f, initial point x^0, initial gap $F_0 := F(x^0)$, desired accuracy $\varepsilon \in (0, F_0]$, inner method $\mathcal{A}$, and phase-length rule $\mathcal{N}$

$z_0 \leftarrow x^0$, $S \leftarrow \lceil \log_2(F_0/\varepsilon) \rceil$
for $s = 0, \dots, S-1$ **do**
  $\Delta_s \leftarrow F_0/2^s$, $N_s \leftarrow \mathcal{N}(\Delta_s)$
  $z_{s+1} \leftarrow \mathcal{A}(f, z_s, N_s)$
end for

Return: z_S

Here $F(z) := f(z) - f^*$ denotes the optimality gap and $\varepsilon \in (0, F_0]$ is the desired accuracy, meaning that the goal is to return z satisfying $F(z) \le \varepsilon$. The point z_s is the starting point of phase s, and $\Delta_s := F_0/2^s$ is the certified gap level at that phase. Thus the certified gap levels form a geometric sequence with ratio $1/2$, so each phase halves the preceding level. The rule $\mathcal{N}(\Delta_s)$ specifies the number N_s of inner iterations needed at this level, while $\mathcal{A}(f, z_s, N_s)$ denotes the corresponding inner method initialized at z_s. The number $S := \lceil \log_2(F_0/\varepsilon) \rceil$ is the total number of phases. The restart invariant is $F(z_s) \le \Delta_s$: each call to $\mathcal{A}$ is chosen to guarantee $F(z_{s+1}) \le \Delta_{s+1} = \Delta_s/2$, and hence $F(z_S) \le \varepsilon$.

Algorithm 1 is reminiscent of the recursive regularization meta-algorithm of Foster et al. (2019): both separate a geometric outer schedule from an oracle-specific inner solver. The mechanisms, however, are different. Foster et al. (2019) change the objective across stages by adding quadratic regularizers with geometrically increasing weights, whereas Algorithm 1 keeps the objective fixed and geometrically decreases the certified upper bound Δ_s on the optimality gap $F(z_s)$. Our construction is therefore a restart scheme that exploits the decreasing effective curvature $H_0 + H_1\Delta_s$, rather than a recursive regularization procedure.

4.1 Full-Gradient Setup

Algorithm 2 Accelerated Full-Gradient Inner Method

Input: objective f, phase start z, iteration budget N, parameters $\theta \in (0, 1]$, $\tau \ge 1$, and $C_\star := 19 + 12\sqrt{2}$

$x_0 \leftarrow z$, $v_0 \leftarrow z$, $A_0 \leftarrow 0$
for $k = 0, \dots, N-1$ **do**
  $y_k \in \arg\min_{\beta \in [0,1]} f\big(v_k + \beta(x_k - v_k)\big)$
  $\widehat{M}_k \leftarrow \max\{2(H_0 + H_1 F(y_k)), \|\nabla f(y_k)\|/r_1\}$
  Choose $M_k \in [\widehat{M}_k, \tau C_\star(H_0 + H_1 F(y_k))]$
  Choose $a_{k+1} > 0$ such that $M_k a_{k+1}^2 = \theta(A_k + a_{k+1})$
  $A_{k+1} \leftarrow A_k + a_{k+1}$
  $x_{k+1} \leftarrow y_k - M_k^{-1}\nabla f(y_k)$
  $v_{k+1} \leftarrow v_k - a_{k+1}\nabla f(y_k)$
end for

Return: x_N

In this subsection, we instantiate the inner method $\mathcal{A}$ from Algorithm 1 in the full-gradient oracle setting, where the algorithm can query $\nabla f(x)$ at each iterate. Our construction builds on the Accelerated Gradient Method with Small-Dimensional Relaxation of Nesterov et al. (2021), which was subsequently extended to convex (L_0, L_1)-smooth objectives by Vankov et al. (2025). The resulting full-gradient inner method is presented in Algorithm 2.

Here z and N are the phase start and inner iteration budget, respectively; x_k is the primal sequence, v_k is the auxiliary accelerated sequence, and A_k is the cumulative weight. The point y_k is obtained by exact one-dimensional minimization over the segment joining v_k and x_k. The two terms in $\widehat{M}_k$ serve different roles: the first controls the local curvature, whereas the second guarantees $\|x_{k+1} - y_k\| = \|\nabla f(y_k)\| / M_k \le r_1$, as required by the radius condition in Assumption 2. The parameter a_{k+1} couples A_{k+1} to the accepted curvature M_k; the x_{k+1}-update is a gradient step, while the v_{k+1}-update produces acceleration. The method returns x_N to the restart wrapper.

Algorithm 2 retains the one-dimensional relaxation and accelerated coupling of the AGMsDR scheme used by Vankov et al. (2025). The main difference is the choice of the effective curvature. In their method, M_k is recovered after the trial update as $M_k := \|\nabla f(y_k)\|^2 / [2(f(y_k) - f(x_{k+1}))]$. Algorithm 2 instead chooses M_k before the update so that it dominates both the gap-dependent curvature and $\|\nabla f(y_k)\| / r_1$; the latter enforces the radius condition in Assumption 2. Moreover, our method is run for a fixed budget N as an inner routine of the restart wrapper.

We now state the convergence guarantee for Algorithm 1 via the full-gradient method in Algorithm 2.

Theorem 5 (Full-gradient complexity). *Suppose Assumptions 1 and 2 hold, and let $\varepsilon \in (0, F_0]$. Define the phase-length rule $\mathcal{N}(\Delta) := \left\lceil 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta} + H_1} \right\rceil$. Then Algorithm 1, with Algorithm 2 as its inner method, returns a point z_S satisfying $F(z_S) \le \varepsilon$. Moreover, the total number of iterations satisfies*

$$N_{\mathrm{tot}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta \varepsilon}} + \widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}} \log \frac{F_0}{\varepsilon} \right).$$

Here $\widetilde{R} := \sup_{x: f(x) \le f(x^0)} \|x - x^*\|$ denotes the finite radius of the initial sublevel set, and $N_{\mathrm{tot}} := \sum_{s=0}^{S-1} N_s$ counts accelerated iterations across all restart phases. The analysis treats the exact segment solve in line 3 of Algorithm 2 as one call per iteration. For fixed θ and τ, Theorem 5 gives $N_{\mathrm{tot}} = O\big(\sqrt{H_0 \widetilde{R}^2/\varepsilon} + \sqrt{H_1 \widetilde{R}^2} \log(F_0/\varepsilon)\big)$. With the implementable inexact line search discussed in Section 6 and a one-dimensional solver whose cost is logarithmic in accuracy, the same bound holds for the total number of oracle calls, including line-search calls, with O replaced by $\widetilde{O}$. When $H_1 = 0$, the first term matches the classical Nesterov dependence (Nesterov, 1983); the second replaces the linear $H_1 R^2$ dependence of non-accelerated analyses by the square-root dependence $\sqrt{H_1 \widetilde{R}^2}$, with only logarithmic dependence on accuracy (Lobanov & Koloskova, 2026). These comparisons concern the dependence on smoothness and accuracy: our bound uses $\widetilde{R}$, whereas the cited bounds use $R := \|x^0 - x^*\|$. The proof is in Appendix C.

Proof sketch. Consider phase s, initialized at z_s with $F(z_s) \le \Delta_s$. The curvature choice keeps each gradient step within the local radius and satisfies $M_k \le \tau C_\star (H_0 + H_1 F(y_k))$; exact segment minimization and the local model give $f(x_{k+1}) \le f(y_k) \le f(x_k)$, hence $F(y_k) \le \Delta_s$. The accelerated potential and weight growth yield $A_{N_s} F(z_{s+1}) \le \frac{1}{2}\|z_s - x^*\|^2$ and $A_{N_s} \ge \theta N_s^2 / [4\tau C_\star (H_0 + H_1 \Delta_s)]$. Together with $\|z_s - x^*\| \le \widetilde{R}$ and $N_s = \mathcal{N}(\Delta_s)$, these bounds give $F(z_{s+1}) \le 2\tau C_\star \widetilde{R}^2 (H_0 + H_1 \Delta_s)/(\theta N_s^2) \le \Delta_s/2$. Hence $F(z_s) \le \Delta_s$ holds by induction, and after $S = \lceil \log_2(F_0/\varepsilon) \rceil$ phases we obtain $F(z_S) \le \varepsilon$. Finally, summing the phase lengths and using $\sum_{s=0}^{S-1} \Delta_s^{-1/2} = O(\varepsilon^{-1/2})$ together with $S = O(\log(F_0/\varepsilon))$ yields the stated iteration complexity.

4.2 Coordinate Setup

Forming a full gradient may be prohibitively expensive in high-dimensional problems. We therefore instantiate the inner method $\mathcal{A}$ using function values and one coordinate derivative $\nabla_i f(x)$ per

iteration, which can reduce the oracle cost whenever a partial derivative is substantially cheaper to evaluate than the full gradient. Such savings are possible even for dense data: for structured problems, Nesterov & Stich (2017) show that both the coordinate-oracle cost and the remaining algorithmic operations can scale linearly with the problem dimension; on such instances, their accelerated coordinate method can outperform standard fast gradient methods. Line-search function differences are likewise inexpensive for models such as quadratic loss and logistic regression (Fountoulakis & Tappenden, 2018).

The resulting coordinate inner method is presented in Algorithm 3. It retains the exact one-dimensional minimization over the segment joining v_k and x_k, as well as the accelerated weight construction of Algorithm 2, but replaces each full-gradient update by a uniformly sampled coordinate update. It is also related to the accelerated coordinate method of Nesterov & Stich (2017): both use one random partial derivative to update the primal and auxiliary sequences. Their method is based on fixed coordinate smoothness constants, whereas ours uses the current gap-dependent curvature, explicitly respects the local-model radius, and is placed inside the restart wrapper. To stay within the page limit, we focus here on uniform sampling and give the non-uniform extension in Appendix A.

Algorithm 3 Accelerated Coordinate Inner Method

Input: objective f, phase start z, iteration budget N, parameters $\theta \in (0,1]$, $\tau \geq 1$, and $C_\star := 19 + 12\sqrt{2}$

1: $x_0 \leftarrow z$, $v_0 \leftarrow z$, $A_0 \leftarrow 0$
2: **for** $k = 0, \dots, N-1$ **do**
3: $\quad y_k \in \arg\min_{\beta \in [0,1]} f\big(v_k + \beta(x_k - v_k)\big)$
4: $\quad M_k \leftarrow \tau C_\star (H_0 + H_1 F(y_k))$
5: $\quad$ Choose $a_{k+1} > 0$ s.t. $d^2 M_k a_{k+1}^2 = \theta(A_k + a_{k+1})$
6: $\quad A_{k+1} \leftarrow A_k + a_{k+1}$
7: $\quad$ Sample $i_k \sim \mathrm{Unif}\{1, \dots, d\}$
8: $\quad x_{k+1} \leftarrow y_k - M_k^{-1} \nabla_{i_k} f(y_k) e_{i_k}$
9: $\quad v_{k+1} \leftarrow v_k - d a_{k+1} \nabla_{i_k} f(y_k) e_{i_k}$
10: **end for**

Return: x_N

Here z, N, x_k, v_k, A_k, and y_k have the same roles as in Algorithm 2; e_i is the i-th standard basis vector, and i_k is sampled uniformly. The single queried derivative $\nabla_{i_k} f(y_k)$ is reused in both updates. Unlike the full-gradient method, the coordinate method does not form $\widehat{M}_k$ from $\|\nabla f(y_k)\|$. Using the gradient bound implied by Assumption 2, it fixes M_k and a_{k+1} before sampling, so $|\nabla_i f(y_k)|/M_k \leq r_1$ for all i and both are independent of i_k. The factors d and d^2 account for uniform sampling, while τ and θ control curvature inflation and accelerated coupling, respectively. The method returns x_N to the restart wrapper.

We now state the convergence guarantee for Algorithm 1 with the coordinate inner method.

Theorem 6 (Coordinate complexity)**.** *Suppose Assumptions 1 and 2 hold, and let $\varepsilon \in (0, F_0]$. Define $\mathcal{N}(\Delta) := \left\lceil 2d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta} + H_1} \right\rceil$. Then Algorithm 1, with Algorithm 3 as its inner method, returns a point z_S satisfying $\mathbb{E}[f(z_S) - f(x^*)] \leq \varepsilon$. Moreover, the total number of iterations satisfies*

$$N_{\mathrm{tot}}^{\mathrm{coord}} = O\left(d\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta \varepsilon}} + d\widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}} \log \frac{F_0}{\varepsilon} \right).$$

For fixed θ and τ, Theorem 6 gives $N_{\mathrm{tot}}^{\mathrm{coord}} = O\big(d\sqrt{H_0 \widetilde{R}^2/\varepsilon} + d\sqrt{H_1 \widetilde{R}^2} \log(F_0/\varepsilon)\big)$, where $N_{\mathrm{tot}}^{\mathrm{coord}}$ counts coordinate-update iterations across all restart phases. Each iteration uses one partial derivative for the main updates and treats the exact segment solve in line 3 of Algorithm 3 as one call. With the inexact line search discussed in Section 6 and logarithmic solver cost, the same bound holds for the total number of oracle calls, including line-search calls, with O replaced by $\widetilde{O}$. Relative to

Theorem 5, the coordinate rate pays the standard factor d; both bounds use the same level-set radius $\widetilde{R}$. When $H_1 = 0$, its dependence matches the classical uniform-sampling accelerated coordinate rate (Nesterov, 2012) and agrees with Nesterov & Stich (2017) when all coordinate smoothness constants coincide. See Appendix D for the proof.

To the best of our knowledge, Theorem 6 is the first accelerated coordinate guarantee for the general class of convex (H_0, H_1)-smooth objectives. Recent coordinate analyses under related non-uniform smoothness conditions concern normalized Gauss–Southwell updates and special losses (Vaswani et al., 2026), rather than accelerated complexity for this general convex class.

Proof sketch. Let $\mathcal{F}_k$ be the σ-algebra generated by all randomness revealed before sampling i_k. Uniform sampling gives $\mathbb{E}[d\nabla_{i_k} f(y_k) e_{i_k} \mid \mathcal{F}_k] = \nabla f(y_k)$ and $\mathbb{E}[(\nabla_{i_k} f(y_k))^2 \mid \mathcal{F}_k] = \|\nabla f(y_k)\|^2 / d$. Since M_k and a_{k+1} are fixed before sampling, these identities combine with the local descent inequality and $d^2 M_k a_{k+1}^2 = \theta A_{k+1}$ to give $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k$, where $\Phi_k := A_k F(x_k) + \frac{1}{2}\|v_k - x^*\|^2$. For a phase starting at z_s, this gives $\mathbb{E}[F(z_{s+1}) \mid z_s] \le 2d^2 \tau C_\star \|z_s - x^*\|^2 (H_0 + H_1 F(z_s))/(\theta N_s^2)$. Using $\|z_s - x^*\| \le \widetilde{R}$ and the induction hypothesis $\mathbb{E}F(z_s) \le \Delta_s$, the choice $N_s = \mathcal{N}(\Delta_s)$ reduces the expected gap by one half. The geometric restart argument and summation of the phase lengths then proceed as in the full-gradient proof, with the additional factor d.

5 Numerical Experiments

In this section, we verify our theoretical results through numerical experiments and report the relative gap $F(x)/F_0$.

Full-gradient algorithms. For $d = 32$, we consider $\min_{x \in \mathbb{R}^d} f_{H_1}(x)$, where $f_{H_1}(x) := \frac{\mu_q}{8} x^\top Q x + \psi_{H_1}(u^\top x)$, $Q := \mathrm{tridiag}(-1, 2, -1)$, $\|u\| = 1$, and $\mu_q = \mu_\psi = 1/2$. Inside a prescribed cutoff, we set $\psi_{H_1}(t) := \frac{\mu_\psi}{H_1}(\cosh(\sqrt{H_1} t) - 1)$ and use its convex C^2 quadratic continuation outside. Because $\|Q\| < 4$, $\psi''_{H_1} = \mu_\psi + H_1 \psi_{H_1}$, and the continuation preserves this bound, $\left\|\nabla^2 f_{H_1}(x)\right\| \le 1 + H_1 F(x)$; thus $H_0 = 1$ and $x^* = 0$, $f^* = 0$. We compare *Ours (restarted AGD)* (Algorithms 1–2) with *GD-Warmup*, the non-accelerated (H_0, H_1)-smooth method of Lobanov & Koloskova (2026); *Frozen-FGM*, the standard two-sequence Nesterov method (Nesterov, 1983), run without restarts or segment relaxation and with the fixed step size $1/M_{\mathrm{frozen}}$, where $M_{\mathrm{frozen}} := H_0 + H_1 F_0$; and *AGMsDR*, Algorithm 1 of Vankov et al. (2025). We choose the continuation so that $L_{\mathrm{global}} := \sup_x \left\|\nabla^2 f_{H_1}(x)\right\| \le M_{\mathrm{frozen}}$ and run AGMsDR with $(L_0, L_1) = (L_{\mathrm{global}}, 0)$. All methods share x^0; we report the first iteration attaining the prescribed relative gap. Figure 1(a) confirms the predicted acceleration: as the gap decreases from 10^{-2} to 10^{-8} at $H_1 = 4$, our count grows from 15 to 105, versus 44 to 2168 for GD-Warmup. In panel (b), our gap-adaptive curvature is less sensitive to H_1: at $H_1 = 64$, it requires 126 iterations versus 1599 for Frozen-FGM. Ours also improves on AGMsDR on this family (105 vs. 136 at gap 10^{-8}, and 126 vs. 429 at $H_1 = 64$), but this does not imply uniform dominance: AGMsDR targets the narrower (L_0, L_1)-smooth class, and Appendix F.1 gives an instance where it needs fewer outer iterations.

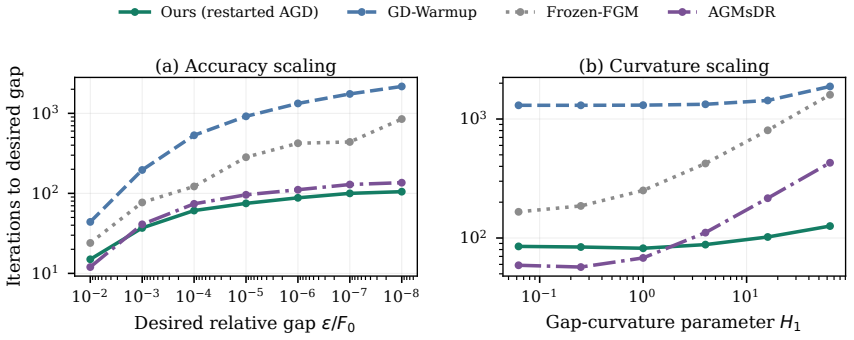

Figure 1: Iterations required on the chain-cosh objective: (a) desired gap for $H_1 = 4$; (b) H_1 for $F(x)/F_0 \le 10^{-6}$.

Coordinate algorithms. We use a 24-dimensional separable capped-cosh objective with four high-curvature coordinates; the remaining coordinates have both $H_{0,i}$ and $H_{1,i}$ reduced by a factor 100. Over 20 seeds and at $F(x)/F_0 \le 10^{-5}$, non-uniform sampling reduces the median coordinate-gradient count from 8170 to 1598 and the total first-order count from 18038 to 4505, supporting the dependence on $S_{1/2}^{(j)} = \sum_i \sqrt{H_{j,i}}$ in Theorem 8. Due to space constraints, Figure 3, which compares our uniform and non-uniform methods, is deferred to Appendix F.2.

6 Discussion and Limitations

Table 1 is representative rather than exhaustive and focuses on rates stated directly in the native (H_0, H_1) parametrization and on their full-gradient or coordinate counterparts. We discuss Tyurin (2025)[1] and Accelerated GRAAL (Borodich & Kovalev, 2026) here rather than adding them to the table: both are important standard first-order full-gradient results for gradient-dependent generalized smoothness, including the (L_0, L_1)-smooth class, but their assumptions, native parameters, and oracle models are not directly comparable with our strict $(H_0, H_1) \setminus (L_0, L_1)$ examples, and neither work develops the coordinate extensions considered here. The closest predecessor is AGMsDR (Vankov et al., 2025), since it uses the same small-dimensional segment relaxation and a closely related accelerated coupling. This is why our main comparison emphasizes Vankov et al.; it does not imply uniform dominance over their method or over the more recent standard first-order algorithms.

One-dimensional line searches are common in first-order optimization, including full-gradient methods; for example, steepest descent may choose its step size by searching along the negative-gradient direction (Nocedal & Wright, 2006; Bottou et al., 2018). The segment relaxation in line 3 of Algorithms 2–4 is related but distinct: it selects a coupling point on $[v_k, x_k]$ before the gradient or coordinate update, rather than its step size (Nesterov et al., 2021; Vankov et al., 2025). The main analysis assumes exact segment minimization. Appendix G.1 replaces it by a verifiable first-order residual, preserving all three iteration bounds up to absolute constants while accounting for line-search queries separately; the resulting cost–accuracy tradeoff is evaluated in Appendix F.3.

All proposed methods require the exact value f^*, because the gap enters the local curvature, the non-uniform sampling probabilities, and the restart test. The value is known in some nonnegative realizable or interpolating models; for example, exactly realizable unregularized squared loss has $f^* = 0$. This observation does not extend to classification without qualification: separable unregularized logistic risk has infimum zero, but that infimum is generally not attained and therefore lies outside our assumption $X^* \neq \varnothing$. With regularization, noise, or model mismatch, f^* is generally unknown, and its removal remains a limitation.

The phase lengths also depend on $\widetilde{R}$, which may be difficult to estimate and much larger than $\left\|x^0 - x^*\right\|$. Appendix G.2 removes it from the inputs by verified doubling, changing the full-gradient cost only by a constant and preserving the coordinate bounds in expectation. Thus $\widetilde{R}$ remains only in the complexity bounds, while valid smoothness bounds are still required.

7 Conclusion

We developed a common acceleration mechanism for convex (H_0, H_1)-smooth optimization and instantiated it with full-gradient, uniform-coordinate, and non-uniform-coordinate methods. The resulting iteration bounds separate the H_0 and H_1 contributions, match the classical smoothness and accuracy dependence when $H_1 = 0$, and exploit coordinate heterogeneity through non-uniform sampling. Acceleration requires only local upper models within their validity radii, rather than a global quadratic model. We complement these guarantees with implementable inexact segment relaxation, whose line-search calls are accounted for separately, and verified doubling that removes prior knowledge of $\widetilde{R}$ from the algorithm inputs. Experiments confirm acceleration, non-uniform sampling gains, and the viability of inexact relaxation. Deriving matching lower bounds and developing stochastic variants remain open directions.

[1]The main text focuses on comparisons with (L_0, L_1)-smooth results; general gradient-dependent ℓ-smoothness is discussed in Appendix H.

AI USE STATEMENT

Generative AI tools were used for language editing and to improve the clarity of the manuscript. We reviewed all resulting changes and take full responsibility for the final content.

REFERENCES

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective. *arXiv preprint arXiv:2510.03164*, 2025.

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtarik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Ekaterina Borodich and Dmitry Kovalev. Nesterov finds graal: Optimal and adaptive gradient method for convex optimization. In *International Conference on Learning Representations*, volume 2026, pp. 95533–95558, 2026.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, 2023.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *Advances in Neural Information Processing Systems*, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Olivier Fercoq and Peter Richtarik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1319–1345. PMLR, 2019.

Kimon Fountoulakis and Rachael Tappenden. A flexible coordinate descent method. *Computational Optimization and Applications*, 70(2):351–394, 2018.

Ofir Gaash, Kfir Y Levy, and Yair Carmon. Convergence of clipped sgd on convex (l_0, l_1)-smooth functions. *Advances in Neural Information Processing Systems*, 38:133483–133515, 2025.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, volume 2025, pp. 97261–97305, 2025.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023a.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023b.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

Yuxing Liu, Yuze Ge, Rui Pan, An Kang, and Tong Zhang. Theoretical analysis on how learning rate warmup accelerates convergence. *arXiv preprint arXiv:2509.07972*, 2025b.

Aleksandr Lobanov and Alexander Gasnikov. Power of generalized smoothness in stochastic convex optimization: First-and zero-order algorithms. *arXiv preprint arXiv:2501.18198*, 2025.

Aleksandr Lobanov and Anastasia Koloskova. Avoiding bias in clipped sgd for overparameterized models under generalized smoothness. *arXiv preprint arXiv:2605.14800*, 2026.

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under (l_0, l_1)-smoothness. *arXiv preprint arXiv:2412.17050*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2018.

Yurii Nesterov and Sebastian U Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1310–1318, 2013.

Peter Richtarik and Martin Takac. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1–2):1–38, 2014.

Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Parameter-free clipped gradient descent meets polyak. *Advances in Neural Information Processing Systems*, 37:44575–44599, 2024.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. *arXiv preprint arXiv:2412.11773*, 2024.

Alexander Tyurin. Near-optimal convergence of accelerated gradient methods under generalized and (l_0, l_1)-smoothness. *arXiv preprint arXiv:2508.06884*, 2025.

Daniil Vankov, Anton Rodomanov, Angelia Nedic, Lalitha Sankar, and Sebastian U. Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sharan Vaswani and Reza Babanezhad. Armijo line-search can make (stochastic) gradient descent provably faster. *arXiv preprint arXiv:2503.00229*, 2025.

Sharan Vaswani, Yifan Sun, and Reza Babanezhad. Convergence of steepest descent and adam under non-uniform smoothness. *arXiv preprint arXiv:2605.30648*, 2026.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Conference on Learning Theory*, 2023.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020b.

APPENDIX

A Few Accelerated Algorithms for Convex Optimization under (H_0, H_1)-Smoothness

CONTENTS

1 Introduction **1**

2 Related Work and Main Contributions **3**

3 Problem setup **4**

4 Accelerated Algorithms **5**

4.1 Full-Gradient Setup 5

4.2 Coordinate Setup 6

5 Numerical experiments **8**

6 Discussion and Limitations **9**

7 Conclusion **9**

A Extension to Non-Uniform Sampling **14**

B Proofs for Examples 3 and 4 **15**

C Proof of Theorem 5 **18**

D Proof of Theorem 6 **24**

E Proof of Theorem 8 **30**

F Additional Experiments **36**

F.1 Full-gradient finite-sum classification 36

F.2 Uniform versus non-uniform coordinate sampling 37

F.3 Inexact Segment-Relaxation Sweep 38

G Practical Extensions **41**

G.1 Inexact Segment Relaxation 41

G.2 Adaptive Phase Lengths without Prior Knowledge of $\widetilde{R}$ 46

H Comparison with General Gradient-Dependent ℓ-Smoothness **49**

H.1 The existing separation examples are ℓ-smooth 49

H.2 Incomparability of the general classes 50

A Extension to Non-Uniform Sampling

Uniform sampling ignores heterogeneity across coordinates. Under classical coordinate smoothness, sampling coordinate i proportionally to the square root of its smoothness constant can replace a worst-coordinate dependence by a sum of coordinate contributions (Nesterov & Stich, 2017; Allen-Zhu et al., 2016). Under (H_0, H_1)-smoothness, both the baseline and gap-dependent curvature may be heterogeneous. We now adapt this importance-sampling principle to that setting.

Assumption 7 (Coordinate (H_0, H_1)-smoothness)**.** *For every $i \in \{1, \dots, d\}$, there exist constants $H_{0,i} \geq 0$ and $H_{1,i} > 0$. Let $\mathcal{L}_i(\Delta) := 2(H_{0,i} + H_{1,i}\Delta)$ and $r_{1,i} := 1/((2\sqrt{3} + \sqrt{6})\sqrt{H_{1,i}})$. For every $x \in \mathbb{R}^d$ and $h \in \mathbb{R}$, if $|h| \leq r_{1,i}$, then*

$$f(x + he_i) \leq f(x) + h\nabla_i f(x) + \frac{\mathcal{L}_i(F(x))}{2}h^2.$$

The case $H_{1,i} = 0$ is covered by $\mathcal{L}_i(\Delta) = 2H_{0,i}$ and $r_{1,i} = \infty$. Assumption 2 implies Assumption 7 with $H_{j,i} = H_j$ for all i and $j \in \{0, 1\}$; the coordinate condition is imposed only along $x + he_i$ and allows coordinate-specific constants and radii. Let $S_{1/2}^{(0)} := \sum_{i=1}^d H_{0,i}^{1/2}$ and $S_{1/2}^{(1)} := \sum_{i=1}^d H_{1,i}^{1/2}$. If $H_{j,i} \leq \rho_i H_j$ for $j \in \{0, 1\}$ and $\rho_i \in [0, 1]$, then $S_{1/2}^{(j)} \leq \sqrt{H_j}\sum_i \sqrt{\rho_i}$. Hence, if curvature is concentrated on $m \ll d$ coordinates and $\sum_i \sqrt{\rho_i} = O(m)$, non-uniform sampling replaces the uniform factor d by m, as in the classical heterogeneous-smoothness setting (Nesterov & Stich, 2017; Allen-Zhu et al., 2016).

This assumption differs from the coordinate (L_0, L_1)-smoothness condition used by Lobanov et al. (2024, Assumption 1.3), which bounds $|\nabla_i f(x+he_i) - \nabla_i f(x)|$ by $(L_{0,i} + L_{1,i}|\nabla_i f(x)|)|h|$ inside a common coordinate radius. Their condition is gradient-dependent, whereas Assumption 7 is gap-dependent and stated directly as the local upper model required by the algorithm. Both recover standard coordinate smoothness when the growth constants vanish.

We assume known valid upper bounds on $\{H_{0,i}, H_{1,i}\}_{i=1}^d$, as in classical non-uniform accelerated coordinate methods (Nesterov & Stich, 2017; Allen-Zhu et al., 2016). Applying the local model at the coordinate-radius boundary gives $|\nabla_i f(x)|/r_{1,i} \leq C_\star(H_{0,i} + H_{1,i}F(x))$ for every i, where $C_\star := 19 + 12\sqrt{2}$; the radius condition is automatic when $H_{1,i} = 0$. This leads to Algorithm 4.

Algorithm 4 Non-Uniform Coordinate Inner Method

Input: objective f, phase start z, iteration budget N, parameters $\theta \in (0, 1]$, $\tau \geq 1$, and $C_\star := 19 + 12\sqrt{2}$

$x_0 \leftarrow z$, $v_0 \leftarrow z$, $A_0 \leftarrow 0$
for $k = 0, \dots, N-1$ **do**
  $y_k \in \arg\min_{\beta \in [0,1]} f\big(v_k + \beta(x_k - v_k)\big)$
  $M_{k,i} \leftarrow \tau C_\star(H_{0,i} + H_{1,i}F(y_k))$, $i = 1, \dots, d$
  $\mathcal{M}_k \leftarrow \sum_{i=1}^d \sqrt{M_{k,i}}$, $p_{k,i} \leftarrow \sqrt{M_{k,i}}/\mathcal{M}_k$
  Choose $a_{k+1} > 0$ s.t. $\mathcal{M}_k^2 a_{k+1}^2 = \theta(A_k + a_{k+1})$
  $A_{k+1} \leftarrow A_k + a_{k+1}$
  Sample $i_k \sim p_k$
  $x_{k+1} \leftarrow y_k - M_{k,i_k}^{-1}\nabla_{i_k} f(y_k)e_{i_k}$
  $v_{k+1} \leftarrow v_k - a_{k+1}p_{k,i_k}^{-1}\nabla_{i_k} f(y_k)e_{i_k}$
end for
Return: x_N

Here $M_{k,i}$ is the local curvature assigned to coordinate i, $\mathcal{M}_k$ normalizes the probabilities, and all $M_{k,i}$, $p_{k,i}$, and a_{k+1} are fixed before sampling i_k. The choice $p_{k,i} \propto \sqrt{M_{k,i}}$ balances coordinate descent with the variance of the auxiliary update. If $H_{0,i} = H_0$ and $H_{1,i} = H_1$ for all i, then $p_{k,i} = 1/d$, $\mathcal{M}_k = d\sqrt{M_k}$, and Algorithm 4 reduces to Algorithm 3. When all $H_{1,i} = 0$, the probabilities become proportional to $\sqrt{H_{0,i}}$, matching the classical square-root importance sampling of Nesterov

& Stich (2017). Unlike their fixed probabilities, however, our probabilities adapt to the current gap through $M_{k,i}$.

We next state the non-uniform coordinate guarantee.

Theorem 8 (Non-uniform coordinate complexity)**.** *Suppose Assumptions 1 and 7 hold, and let $\varepsilon \in (0, F_0]$. Define $\mathcal{N}(\Delta) := \left\lceil 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}(S_{1/2}^{(0)}/\sqrt{\Delta} + S_{1/2}^{(1)}) \right\rceil$. Then Algorithm 1, with Algorithm 4 as its inner method, returns a point z_S satisfying $\mathbb{E}[f(z_S) - f(x^*)] \le \varepsilon$. Moreover, the total number of coordinate-update iterations satisfies*

$$N_{\mathrm{tot}}^{\mathrm{nonunif}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\left(\frac{S_{1/2}^{(0)}}{\sqrt{\varepsilon}} + S_{1/2}^{(1)} \log \frac{F_0}{\varepsilon}\right)\right).$$

For fixed θ and τ, Theorem 8 gives $N_{\mathrm{tot}}^{\mathrm{nonunif}} = O(\widetilde{R}S_{1/2}^{(0)}/\sqrt{\varepsilon} + \widetilde{R}S_{1/2}^{(1)} \log(F_0/\varepsilon))$, where $N_{\mathrm{tot}}^{\mathrm{nonunif}}$ counts importance-sampled coordinate-update iterations across all restart phases. Each iteration uses one sampled partial derivative for the main updates and treats the exact segment solve in line 3 of Algorithm 4 as one call. With the inexact line search discussed in Section 6 and logarithmic solver cost, the same bound holds for the total number of oracle calls, including line-search calls, with O replaced by $\widetilde{O}$. In the uniform case, $S_{1/2}^{(0)} = d\sqrt{H_0}$ and $S_{1/2}^{(1)} = d\sqrt{H_1}$, recovering Theorem 6; heterogeneous coordinate curvatures can make both sums substantially smaller. When $H_{1,i} = 0$ for all i, the first term recovers the standard non-uniform accelerated coordinate dependence on $\sum_i \sqrt{H_{0,i}}$ (Nesterov & Stich, 2017; Allen-Zhu et al., 2016). See Appendix E for the proof.

To the best of our knowledge, Theorem 8 is the first accelerated non-uniform coordinate guarantee under coordinate-wise (H_0, H_1)-smoothness. In contrast, the coordinate (L_0, L_1)-smooth analysis of Lobanov et al. (2024) uses uniform sampling and gives a non-accelerated rate.

Proof sketch. The proof follows Theorem 6. Conditioning before sampling i_k, the choices $p_{k,i} = \sqrt{M_{k,i}}/\mathcal{M}_k$ and $\mathcal{M}_k^2 a_{k+1}^2 = \theta A_{k+1}$ preserve the same potential decrease. The uniform quantity $d\sqrt{M_k}$ is replaced by $\mathcal{M}_k$, where $\mathcal{M}_k \le \sqrt{\tau C_\star}(S_{1/2}^{(0)} + S_{1/2}^{(1)}\sqrt{F(y_k)})$. The same monotonicity and restart arguments then yield the stated complexity.

Remark 9 (Phase-wise sampling)**.** Although Algorithm 4 updates the sampling probabilities at every iteration, convergence only requires them to be fixed before sampling i_k. To avoid an $O(d)$ probability update per iteration, one may set $M_{s,i} := \tau C_\star(H_{0,i} + H_{1,i}F(z_s))$ and $p_{s,i} := \sqrt{M_{s,i}}/\sum_j \sqrt{M_{s,j}}$ once at the beginning of phase s. Since $F(y_k) \le F(z_s)$ throughout the phase, the same radius and descent conditions hold, and the complexity remains unchanged.

B Proofs for Examples 3 and 4

Example 3 (One-dimensional). *Fix $p \in (1, 2]$. Let $q_p(t) := 1 + (1 + t^2)^{p/2} \exp(-(1 + t^2)^p \sin^2(\pi t))$, and define*

$$\phi_p(x) = \int_0^x \int_0^s q_p(t)\, dt\, ds.$$

Then ϕ_p is convex and satisfies Assumption 2 with $H_0 = H_1 = 2$, but is not (L_0, L_1)-smooth for any finite L_0, L_1.

Proof. The function q_p is infinitely differentiable because it is a composition of infinitely differentiable functions and $1 + t^2$ is strictly positive. The fundamental theorem of calculus therefore gives $\phi_p'(x) = \int_0^x q_p(t)\, dt$ and $\phi_p''(x) = q_p(x)$. In particular, $\phi_p(0) = \phi_p'(0) = 0$. Since the exponential term in q_p is nonnegative, $q_p(x) \ge 1$ for every x. Taylor's formula with integral remainder

consequently yields

$$\phi_p(x) = x^2 \int_0^1 (1-s)\phi_p''(sx)\,ds = x^2 \int_0^1 (1-s)q_p(sx)\,ds \geq x^2 \int_0^1 (1-s)\,ds = \frac{x^2}{2}. \tag{3}$$

The first equality uses $\phi_p(0) = \phi_p'(0) = 0$, the second uses $\phi_p'' = q_p$, and the inequality uses $q_p \geq 1$. Hence ϕ_p is 1-strongly convex, and its unique minimizer is $x^* = 0$, with $\phi_p^* = \phi_p(0) = 0$.

For every $x \in \mathbb{R}$,

$$\phi_p''(x) = q_p(x) \leq 1 + (1+x^2)^{p/2} \leq 2 + x^2 \leq 2 + 2\phi_p(x). \tag{4}$$

The first inequality follows from $\exp(-u) \leq 1$ for $u \geq 0$. The second uses $p/2 \leq 1$ and $1 + x^2 \geq 1$, which imply $(1+x^2)^{p/2} \leq 1 + x^2$. The last inequality follows from equation 3. Thus $|\phi_p''(x)| \leq 2 + 2(\phi_p(x) - \phi_p^*)$ for all x. Applying the local-model implication of Liu et al. (2025b, Lemma 2) with $\rho = 1$ therefore shows that ϕ_p satisfies Assumption 2 with $H_0 = H_1 = 2$.

Suppose, toward a contradiction, that ϕ_p is (L_0, L_1)-smooth for some finite $L_0, L_1 \geq 0$. Since ϕ_p is twice continuously differentiable, the original Hessian formulation of (L_0, L_1)-smoothness introduced by Zhang et al. (2020b) requires the pointwise condition

$$|\phi_p''(x)| \leq L_0 + L_1|\phi_p'(x)|, \qquad x \in \mathbb{R}. \tag{5}$$

The same necessary condition follows from the first-order local-gradient formulation of Zhang et al. (2020a): divide its inequality by $|y - x|$ and let $y \to x$, so that the difference quotient converges to $|\phi_p''(x)|$. At every integer $n \geq 1$, $\sin(\pi n) = 0$, and hence

$$\phi_p''(n) = q_p(n) = 1 + (1+n^2)^{p/2} \geq n^p, \tag{6}$$

where the inequality follows from $1 + n^2 \geq n^2$.

We next bound $\phi_p'(n)$. Set $a(t) := (1+t^2)^{p/2}$, $I_m := [m - \frac{1}{2}, m + \frac{1}{2}]$, and $A_m := (1+m^2)^{p/2}$. If $m \geq 1$ and $t \in I_m$, then $m/2 \leq t \leq 3m/2$, so

$$\frac{1+m^2}{4} \leq 1 + t^2 \leq \frac{9}{4}(1+m^2).$$

Raising these inequalities to the positive power $p/2$ and using $p \leq 2$ gives

$$\frac{A_m}{4} \leq 2^{-p} A_m \leq a(t) \leq \left(\frac{3}{2}\right)^p A_m \leq \frac{9}{4} A_m. \tag{7}$$

For $u := t - m \in [-\frac{1}{2}, \frac{1}{2}]$, concavity of $\sin(\pi u)$ on $[0, \frac{1}{2}]$, together with symmetry, gives $|\sin(\pi t)| = |\sin(\pi u)| \geq 2|u|$. Combining $|\sin(\pi t)| \geq 2|u|$ with the lower bound on $a(t)$ in equation 7, we obtain $a(t)^2 \sin^2(\pi t) \geq A_m^2 u^2/4$. Therefore,

$$\begin{aligned} &\int_{I_m} a(t) \exp\bigl(-a(t)^2 \sin^2(\pi t)\bigr)\,dt \\ &\quad \leq \frac{9}{4} A_m \int_{-1/2}^{1/2} \exp\left(-\frac{A_m^2 u^2}{4}\right) du \leq \frac{9}{4} A_m \int_{-\infty}^{\infty} \exp\left(-\frac{A_m^2 u^2}{4}\right) du = \frac{9}{2}\sqrt{\pi}. \end{aligned} \tag{8}$$

The first inequality uses the upper bound $a(t) \leq 9A_m/4$ and the lower bound on the exponent, the second enlarges the integration domain, and the last equality is the Gaussian integral after the substitution $v = A_m u/2$. For $m = 0$, $t \in I_0$ implies $a(t) \leq (5/4)^{p/2} \leq 5/4$, so the same integral over I_0 is at most $5/4 < 9\sqrt{\pi}/2$. The intervals $I_0, \ldots, I_n$ cover $[0, n]$ up to endpoints. Since the integrand is nonnegative, equation 8 and the $m = 0$ bound give, for every integer $n \geq 1$,

$$\begin{aligned} \phi_p'(n) &= \int_0^n q_p(t)\,dt \\ &= n + \int_0^n a(t) \exp\bigl(-a(t)^2 \sin^2(\pi t)\bigr)\,dt \\ &\leq n + \sum_{m=0}^{n} \int_{I_m} a(t) \exp\bigl(-a(t)^2 \sin^2(\pi t)\bigr)\,dt \\ &\leq n + \frac{9}{2}\sqrt{\pi}(n+1) \leq (1 + 9\sqrt{\pi})n. \end{aligned} \tag{9}$$

The last inequality uses $n + 1 \le 2n$, valid for $n \ge 1$. Substituting equation 6 and equation 9 into the necessary (L_0, L_1)-smoothness condition equation 5 gives

$$n^p \le L_0 + L_1(1 + 9\sqrt{\pi})n.$$

After division by n, the left-hand side is $n^{p-1} \to \infty$ because $p > 1$, whereas the right-hand side is at most $L_0 + L_1(1 + 9\sqrt{\pi})$. This contradiction proves that no finite pair (L_0, L_1) can satisfy the generalized smoothness condition for ϕ_p. □

Example 4 (Multidimensional). *For $d \ge 2$, let*

$$\Phi_p(x_1, \dots, x_d) = \phi_p(x_1) + \frac{1}{2}\sum_{i=2}^{d} x_i^2,$$

where ϕ_p is the function from Example 3. Then Φ_p is strongly convex and satisfies Assumption 2 with $H_0 = H_1 = 2$, but is not (L_0, L_1)-smooth for any finite L_0, L_1.

Proof. Differentiating the definition of Φ_p gives

$$\nabla\Phi_p(x) = \big(\phi_p'(x_1), x_2, \dots, x_d\big)$$

and

$$\nabla^2\Phi_p(x) = \operatorname{diag}\big(\phi_p''(x_1), 1, \dots, 1\big).$$

The proof of Example 3 shows that $\phi_p''(x_1) \ge 1$. Hence $\nabla^2\Phi_p(x) \succeq I_d$, so Φ_p is 1-strongly convex. Moreover, $\nabla\Phi_p(\mathbf{0}) = \mathbf{0}$ and $\Phi_p(\mathbf{0}) = 0$; therefore $\mathbf{0} \in \mathbb{R}^d$ is the unique minimizer and $\Phi_p^* = 0$.

Since $\phi_p''(x_1) \ge 1$, the largest diagonal entry of the Hessian is $\phi_p''(x_1)$. Consequently,

$$\left\|\nabla^2\Phi_p(x)\right\| = \phi_p''(x_1) \le 2 + 2\phi_p(x_1) \le 2 + 2\Phi_p(x).$$

Here the first inequality follows from equation 4, and the second follows from $\Phi_p(x) = \phi_p(x_1) + \frac{1}{2}\sum_{i=2}^{d} x_i^2 \ge \phi_p(x_1)$. Thus $\left\|\nabla^2\Phi_p(x)\right\| \le 2 + 2(\Phi_p(x) - \Phi_p^*)$, and the local-model implication of Liu et al. (2025b, Lemma 2) with $\rho = 1$ shows that Φ_p satisfies Assumption 2 with $H_0 = H_1 = 2$.

Suppose, toward a contradiction, that Φ_p is (L_0, L_1)-smooth for some finite $L_0, L_1 \ge 0$. Because Φ_p is twice continuously differentiable, the Hessian formulation of Zhang et al. (2020b) requires

$$\left\|\nabla^2\Phi_p(x)\right\| \le L_0 + L_1 \left\|\nabla\Phi_p(x)\right\|, \qquad x \in \mathbb{R}^d. \tag{10}$$

Equivalently, equation 10 follows from the first-order formulation of Zhang et al. (2020a) by taking $y = x + tv$ for an arbitrary unit vector v, dividing by $|t|$, letting $t \to 0$, and then taking the supremum over all such v. For every integer $n \ge 1$, set $x^{(n)} := (n, 0, \dots, 0)$. The estimates in equation 6 and equation 9 give

$$\left\|\nabla^2\Phi_p(x^{(n)})\right\| = \phi_p''(n) \ge n^p$$

and

$$\left\|\nabla\Phi_p(x^{(n)})\right\| = |\phi_p'(n)| \le (1 + 9\sqrt{\pi})n.$$

Substitution into equation 10 yields

$$n^p \le L_0 + L_1(1 + 9\sqrt{\pi})n.$$

Dividing by n gives $n^{p-1} \le L_0/n + L_1(1 + 9\sqrt{\pi})$. The right-hand side remains bounded as $n \to \infty$, whereas the left-hand side diverges because $p > 1$. This contradiction proves that Φ_p is not (L_0, L_1)-smooth for any finite L_0, L_1. □

C Proof of Theorem 5

Before proving Theorem 5, we establish an auxiliary curvature estimate that will also be used in the uniform- and non-uniform-coordinate analyses. It shows that enforcing the local-model radius changes the corresponding curvature only by an absolute factor.

Lemma 10 (Curvature bound for the local radius)**.** *Let $C_\star := 19 + 12\sqrt{2}$. Under Assumption 2, for every $x \in \mathbb{R}^d$,*

$$\max\left\{2\big(H_0 + H_1F(x)\big), \frac{\|\nabla f(x)\|}{r_1}\right\} \le C_\star\big(H_0 + H_1F(x)\big).$$

Under Assumption 7, for every $x \in \mathbb{R}^d$ and $i \in \{1, \dots, d\}$,

$$\max\left\{2\big(H_{0,i} + H_{1,i}F(x)\big), \frac{|\nabla_i f(x)|}{r_{1,i}}\right\} \le C_\star\big(H_{0,i} + H_{1,i}F(x)\big).$$

When $H_1 = 0$ or $H_{1,i} = 0$, respectively, the corresponding ratio with the infinite radius is interpreted as zero.

Proof. We first prove the full-gradient statement. The case $H_1 = 0$ is immediate because $r_1 = \infty$ and $C_\star > 2$. Suppose $H_1 > 0$. If $\nabla f(x) = 0$, the ratio involving the gradient is zero. Otherwise, set

$$\bar{x} := x - r_1 \frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Since $\|\bar{x} - x\| = r_1$, Assumption 2 applies to $\bar{x}$ with base point x. Using $\langle \nabla f(x), \bar{x} - x\rangle = -r_1 \|\nabla f(x)\|$ and $f^* \le f(\bar{x})$, we obtain

$$f^* \le f(x) - r_1 \|\nabla f(x)\| + \big(H_0 + H_1F(x)\big)r_1^2.$$

Subtracting f^*, rearranging, and dividing by r_1^2 gives

$$\begin{aligned}\frac{\|\nabla f(x)\|}{r_1} &\le \frac{F(x)}{r_1^2} + H_0 + H_1F(x)\\ &= H_0 + (19 + 12\sqrt{2})H_1F(x)\\ &\le C_\star\big(H_0 + H_1F(x)\big),\end{aligned}$$

where we used $r_1^{-2} = (2\sqrt{3} + \sqrt{6})^2 H_1 = (18 + 12\sqrt{2})H_1$. Thus the coefficient $19 + 12\sqrt{2}$ is obtained by adding the remaining $H_1F(x)$ term to $(18 + 12\sqrt{2})H_1F(x)$. The other term in the maximum is bounded by the same expression because $C_\star > 2$, proving the first claim.

For the coordinate statement, the case $H_{1,i} = 0$ is again immediate. Suppose $H_{1,i} > 0$ and $\nabla_i f(x) \ne 0$, and set $h := -r_{1,i}\operatorname{sign}(\nabla_i f(x))$. Then $|h| = r_{1,i}$, so Assumption 7 and $f^* \le f(x + he_i)$ imply

$$f^* \le f(x) - r_{1,i}|\nabla_i f(x)| + \big(H_{0,i} + H_{1,i}F(x)\big)r_{1,i}^2.$$

Therefore,

$$\begin{aligned}\frac{|\nabla_i f(x)|}{r_{1,i}} &\le \frac{F(x)}{r_{1,i}^2} + H_{0,i} + H_{1,i}F(x)\\ &= H_{0,i} + (19 + 12\sqrt{2})H_{1,i}F(x)\\ &\le C_\star\big(H_{0,i} + H_{1,i}F(x)\big).\end{aligned}$$

Here $r_{1,i}^{-2} = (18 + 12\sqrt{2})H_{1,i}$, and the additional $H_{1,i}F(x)$ term again increases the coefficient from $18 + 12\sqrt{2}$ to $19 + 12\sqrt{2}$. The inequality is immediate when $\nabla_i f(x) = 0$, and the first term in the maximum is controlled because $C_\star > 2$. This proves the coordinate statement. □

With this auxiliary estimate in place, we now restate the full-gradient result and provide its complete proof.

Theorem 5 (Full-gradient complexity). *Suppose Assumptions 1 and 2 hold, and let* $\varepsilon \in (0, F_0]$. *Define the phase-length rule* $\mathcal{N}(\Delta) := \left\lceil 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta} + H_1} \right\rceil$. *Then Algorithm 1, with Algorithm 2 as its inner method, returns a point* z_S *satisfying* $F(z_S) \le \varepsilon$. *Moreover, the total number of iterations satisfies*

$$N_{\mathrm{tot}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta\varepsilon}} + \widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}}\log\frac{F_0}{\varepsilon}\right).$$

Proof. The proof has five main steps. First, the curvature choice in Algorithm 2 ensures $M_k \ge \|\nabla f(y_k)\|/r_1$, where the local-model radius from Assumption 2 is $r_1 = ((2\sqrt{3}+\sqrt{6})\sqrt{H_1})^{-1}$ for $H_1 > 0$ and $r_1 = \infty$ for $H_1 = 0$. Hence $\|x_{k+1} - y_k\| \le r_1$, and the local model gives equation 12; Lemma 10 shows that enforcing this radius changes the curvature only by the absolute factor $C_\star$, as summarized in equation 13. Second, exact segment minimization and convexity give the coupling inequality equation 15, which leads to the potential bound equation 18. Third, a lower bound on the accelerated weights converts this potential bound into the one-run estimate equation 22. Fourth, the phase-length rule in Theorem 5 turns that estimate into the gap contraction equation 28. Finally, induction over the restart phases and summation of their lengths give the desired accuracy and total complexity.

Local descent and curvature control. We use the first-order local model in Assumption 2 directly. Recall that $r_1 := 1/((2\sqrt{3}+\sqrt{6})\sqrt{H_1})$ for $H_1 > 0$, while $r_1 := \infty$ in the limiting case $H_1 = 0$. Whenever $\|y - x\| \le r_1$,

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \big(H_0 + H_1 F(x)\big)\|y - x\|^2.$$

The radius condition is essential. In the accelerated gradient step

$$x_{k+1} = y_k - \frac{1}{M_k}\nabla f(y_k),$$

it becomes

$$\|x_{k+1} - y_k\| = \frac{\|\nabla f(y_k)\|}{M_k} \le r_1.$$

This is not implied by choosing $M_k = 2(H_0 + H_1 F(y_k))$, and is the reason for the second term in the definition of $\widehat{M}_k$ in Algorithm 2.

At iteration k, Algorithm 2 chooses

$$M_k \ge \widehat{M}_k := \max\left\{2\big(H_0 + H_1 F(y_k)\big), \frac{\|\nabla f(y_k)\|}{r_1}\right\}.$$

Then

$$\|x_{k+1} - y_k\| = \left\|-\frac{1}{M_k}\nabla f(y_k)\right\| = \frac{\|\nabla f(y_k)\|}{M_k} \le r_1.$$

Indeed, the first equality follows by subtracting y_k from the update defining x_{k+1}, and the second follows from the absolute homogeneity of the norm. Moreover, $M_k \ge \widehat{M}_k \ge \|\nabla f(y_k)\|/r_1$, so multiplying the latter inequality by r_1/M_k for $H_1 > 0$ gives $\|\nabla f(y_k)\|/M_k \le r_1$. When $H_1 = 0$, we have $r_1 = \infty$, and the radius condition is vacuous. Thus the local model applies with $x = y_k$ and $y = x_{k+1}$. Therefore

$$f(x_{k+1}) \le f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \big(H_0 + H_1 F(y_k)\big)\|x_{k+1} - y_k\|^2.$$

Since $x_{k+1} - y_k = -M_k^{-1}\nabla f(y_k)$,

$$f(x_{k+1}) \le f(y_k) - \frac{1}{M_k}\|\nabla f(y_k)\|^2 + \frac{H_0 + H_1 F(y_k)}{M_k^2}\|\nabla f(y_k)\|^2. \tag{11}$$

Because $M_k \geq 2(H_0 + H_1 F(y_k))$, we have $(H_0 + H_1 F(y_k))/M_k^2 \leq 1/(2M_k)$. Substituting this inequality into equation 11 gives

$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2M_k} \|\nabla f(y_k)\|^2 . \tag{12}$$

Applying Lemma 10 at y_k gives

$$\widehat{M}_k \leq C_\star(H_0 + H_1 F(y_k)).$$

Since $\tau \geq 1$, the upper endpoint $\tau C_\star(H_0 + H_1 F(y_k))$ is no smaller than $\widehat{M}_k$. Thus the interval used to choose M_k in Algorithm 2 is nonempty, and every accepted curvature satisfies

$$\widehat{M}_k \leq M_k \leq \tau C_\star\big(H_0 + H_1 F(y_k)\big). \tag{13}$$

Finally, since $\beta = 1$ is feasible in the inner segment-relaxation problem, $f(y_k) = \min_{\beta\in[0,1]} f(v_k + \beta(x_k - v_k)) \leq f(x_k)$, while Equation equation 12 gives

$$f(x_{k+1}) \leq f(y_k) \leq f(x_k). \tag{14}$$

Since $x_0 = z$, induction on k gives $f(x_k) \leq f(z)$ and $f(y_k) \leq f(z)$ for every iteration of the inner run. Thus the primal iterates and interpolation points remain in the sublevel set determined by the phase start z; no analogous claim is needed for the auxiliary sequence $\{v_k\}$.

Accelerated coupling and potential. Fix any $x^* \in X^*$, and consider one inner run initialized at $x_0 = v_0 = z$ and $A_0 = 0$. At iteration k, let

$$y_k \in \operatorname*{arg\,min}_{\beta\in[0,1]} f\big(v_k + \beta(x_k - v_k)\big).$$

For any $a > 0$, define $A^+ := A_k + a$ and

$$w_k(a) := \frac{A_k x_k + a v_k}{A^+}.$$

The coefficients in $w_k(a)$ are nonnegative and sum to one, so $w_k(a)$ lies on the segment between x_k and v_k. The first-order optimality condition for the convex minimization defining y_k therefore gives

$$\langle \nabla f(y_k), w_k(a) - y_k \rangle \geq 0.$$

By convexity,

$$f(y_k) - f(x_k) \leq \langle \nabla f(y_k), y_k - x_k \rangle , \qquad F(y_k) \leq \langle \nabla f(y_k), y_k - x^* \rangle .$$

The first inequality is obtained by applying convexity at y_k with the comparison point x_k; the second is obtained with the comparison point x^*, using $f(x^*) = f^*$. Since $A^+ = A_k + a$, these inequalities give

$$\begin{aligned} A^+ F(y_k) - A_k F(x_k) &= A_k\big(f(y_k) - f(x_k)\big) + a F(y_k) \\ &\leq A_k \langle \nabla f(y_k), y_k - x_k \rangle + a \langle \nabla f(y_k), y_k - x^* \rangle . \end{aligned}$$

The vector identity

$$A_k(y_k - x_k) + a(y_k - x^*) = a(v_k - x^*) + A^+(y_k - w_k(a))$$

follows by substituting the definition of $w_k(a)$ and collecting terms. Taking its inner product with $\nabla f(y_k)$, and using $\langle \nabla f(y_k), y_k - w_k(a) \rangle \leq 0$, gives

$$A^+ F(y_k) - A_k F(x_k) \leq a \langle \nabla f(y_k), v_k - x^* \rangle .$$

Taking $a = a_{k+1}$ gives the coupling inequality

$$A_{k+1} F(y_k) - A_k F(x_k) \leq a_{k+1} \langle \nabla f(y_k), v_k - x^* \rangle . \tag{15}$$

Define the potential

$$\Phi_k = A_k F(x_k) + \frac{1}{2} \|v_k - x^*\|^2 .$$

Multiplying equation 12 by the nonnegative weight A_{k+1} gives

$$A_{k+1}F(x_{k+1}) \le A_{k+1}F(y_k) - \frac{A_{k+1}}{2M_k}\left\|\nabla f(y_k)\right\|^2. \tag{16}$$

Expanding the squared norm in the update $v_{k+1} = v_k - a_{k+1}\nabla f(y_k)$ gives the identity

$$\frac{1}{2}\left\|v_{k+1} - x^*\right\|^2 = \frac{1}{2}\left\|v_k - x^*\right\|^2 - a_{k+1}\left\langle \nabla f(y_k), v_k - x^*\right\rangle + \frac{a_{k+1}^2}{2}\left\|\nabla f(y_k)\right\|^2. \tag{17}$$

Adding equation 16 and equation 17, and then using equation 15 to bound $A_{k+1}F(y_k) - a_{k+1}\left\langle \nabla f(y_k), v_k - x^*\right\rangle$ by $A_kF(x_k)$, gives

$$\Phi_{k+1} \le \Phi_k + \left(\frac{a_{k+1}^2}{2} - \frac{A_{k+1}}{2M_k}\right)\left\|\nabla f(y_k)\right\|^2.$$

The defining relation $M_k a_{k+1}^2 = \theta A_{k+1}$ yields

$$\frac{a_{k+1}^2}{2} - \frac{A_{k+1}}{2M_k} = -\frac{(1-\theta)A_{k+1}}{2M_k} \le 0.$$

The last inequality uses $\theta \in (0, 1]$, and hence $\Phi_{k+1} \le \Phi_k$. Repeated application from $k = 0$ to $N - 1$, together with $A_0 = 0$ and $v_0 = z$, yields

$$A_N F(x_N) \le \Phi_N \le \Phi_0 = \frac{1}{2}\left\|z - x^*\right\|^2. \tag{18}$$

Here the first inequality follows because the second term $\frac{1}{2}\left\|v_N - x^*\right\|^2$ in Φ_N is nonnegative.

One-run estimate. We now derive a lower bound on A_N. Since

$$M_k a_{k+1}^2 = \theta A_{k+1}, \qquad A_{k+1} - A_k = a_{k+1},$$

and $a_{k+1} > 0$, the sequence $\{A_k\}$ is nondecreasing. Therefore

$$\sqrt{A_{k+1}} - \sqrt{A_k} = \frac{a_{k+1}}{\sqrt{A_{k+1}} + \sqrt{A_k}} \ge \frac{a_{k+1}}{2\sqrt{A_{k+1}}} = \frac{1}{2}\sqrt{\frac{\theta}{M_k}}.$$

The first equality is the difference-of-squares identity, the inequality uses $A_k \le A_{k+1}$, and the last equality follows from $M_k a_{k+1}^2 = \theta A_{k+1}$. Summing over $k = 0, \ldots, N-1$ telescopes the left-hand side; because $A_0 = 0$, this gives

$$\sqrt{A_N} \ge \frac{\sqrt{\theta}}{2}\sum_{k=0}^{N-1}\frac{1}{\sqrt{M_k}}. \tag{19}$$

The upper bound in equation 13 implies

$$\frac{1}{\sqrt{M_k}} \ge \frac{1}{\sqrt{\tau C_\star}}\frac{1}{\sqrt{H_0 + H_1 F(y_k)}}. \tag{20}$$

Substituting equation 20 into equation 19 and squaring both nonnegative sides yields

$$A_N \ge \frac{\theta}{4\tau C_\star}\left(\sum_{k=0}^{N-1}\frac{1}{\sqrt{H_0 + H_1 F(y_k)}}\right)^2. \tag{21}$$

Combining equation 21 with equation 18, and then dividing by the positive quantity A_N, gives

$$F(x_N) \le \frac{2\tau C_\star \left\|z - x^*\right\|^2}{\theta\left(\sum_{k=0}^{N-1}\frac{1}{\sqrt{H_0 + H_1 F(y_k)}}\right)^2}. \tag{22}$$

Restart radius and phase contraction. We first isolate the geometric fact that makes a uniform restart radius available.

Lemma 11 (Bounded initial sublevel set)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and convex, and suppose that $X^* := \arg\min_{x\in\mathbb{R}^d} f(x)$ is nonempty and bounded. Then, for every $x^0 \in \mathbb{R}^d$, the initial sublevel set $\mathcal{C}_0 := \{x \in \mathbb{R}^d : f(x) \le f(x^0)\}$ is compact. Consequently, for any fixed $x^* \in X^*$,*

$$\widetilde{R} := \sup_{f(x)\le f(x^0)} \|x - x^*\| \tag{23}$$

is finite.

Proof of Lemma 11. Continuity of f implies that $\mathcal{C}_0$ is closed. Suppose for contradiction that $\mathcal{C}_0$ is unbounded. Then there is a sequence $\{u_j\} \subset \mathcal{C}_0$ such that $r_j := \|u_j - x^*\| \to \infty$. Define $d_j := (u_j - x^*)/r_j$. Every d_j has unit norm, so compactness of the unit sphere gives a subsequence, denoted in the same way, such that $d_j \to d$ for some d with $\|d\| = 1$. Fix any $t \ge 0$. For all sufficiently large j, $t \le r_j$, and

$$x^* + td_j = \left(1 - \frac{t}{r_j}\right) x^* + \frac{t}{r_j} u_j.$$

The coefficients on the right are nonnegative and sum to one. Convexity and the inequalities $f(x^*) = f^*$ and $f(u_j) \le f(x^0)$ therefore imply

$$f(x^* + td_j) \le \left(1 - \frac{t}{r_j}\right) f^* + \frac{t}{r_j} f(x^0) = f^* + \frac{t}{r_j} F_0.$$

Letting $j \to \infty$ and using continuity gives $f(x^* + td) \le f^*$. Since f^* is the minimum value, equality must hold, and hence $x^* + td \in X^*$ for every $t \ge 0$. This produces an unbounded ray in X^*, contradicting the assumed boundedness of the solution set. Thus $\mathcal{C}_0$ is bounded and, being closed, compact. The continuous map $x \mapsto \|x - x^*\|$ attains a finite maximum on $\mathcal{C}_0$, proving the final claim. □

Let phase s start from z_s with $F(z_s) \le \Delta_s$, and let the inner method run for N_s iterations. By equation 14 and $x_0 = z_s$, $f(y_k) \le f(x_k) \le f(z_s)$, so every interpolation point in this phase satisfies

$$F(y_k) \le \Delta_s. \tag{24}$$

Moreover, monotonicity keeps z_s in $\mathcal{C}_0$, so Lemma 11 yields

$$\|z_s - x^*\| \le \widetilde{R}. \tag{25}$$

The bound equation 24, together with $H_1 \ge 0$, implies

$$\sum_{k=0}^{N_s-1} \frac{1}{\sqrt{H_0 + H_1 F(y_k)}} \ge \frac{N_s}{\sqrt{H_0 + H_1\Delta_s}}. \tag{26}$$

Indeed, each denominator on the left is no larger than $\sqrt{H_0 + H_1\Delta_s}$. Applying equation 22 with $z = z_s$, $N = N_s$, and $z_{s+1} = x_{N_s}$, and then using equation 25 and equation 26, gives

$$F(z_{s+1}) \le \frac{2\tau C_\star \widetilde{R}^2}{\theta} \frac{H_0 + H_1\Delta_s}{N_s^2}. \tag{27}$$

The phase-length rule ensures

$$N_s \ge 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta_s} + H_1}.$$

Squaring this inequality gives

$$N_s^2 \ge \frac{4\tau C_\star \widetilde{R}^2}{\theta} \frac{H_0 + H_1\Delta_s}{\Delta_s}.$$

Substitution into the phase estimate equation 27 yields

$$F(z_{s+1}) \le \frac{2\tau C_\star \widetilde{R}^2}{\theta}(H_0 + H_1\Delta_s)\frac{\theta\Delta_s}{4\tau C_\star \widetilde{R}^2(H_0 + H_1\Delta_s)} = \frac{\Delta_s}{2} = \Delta_{s+1}. \tag{28}$$

The equality $\Delta_{s+1} = \Delta_s/2$ follows from $\Delta_s = F_0/2^s$. Since $z_0 = x^0$ and $F(z_0) = F_0 = \Delta_0$, repeated application of equation 28 proves

$$F(z_s) \le \Delta_s.$$

For

$$S = \left\lceil \log_2 \frac{F_0}{\varepsilon} \right\rceil$$

we have $2^S \ge F_0/\varepsilon$, and hence $\Delta_S = F_0/2^S \le \varepsilon$. Combining this with the restart invariant gives $F(z_S) \le \Delta_S \le \varepsilon$.

Total complexity. The total number of inner iterations is

$$N_{\mathrm{tot}} = \sum_{s=0}^{S-1} N_s.$$

For brevity, denote the quantity inside the ceiling in the phase-length rule by

$$T_s := 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta_s} + H_1}, \qquad N_s = \lceil T_s \rceil. \tag{29}$$

We first show that the ceiling contributes only an absolute factor. Since $\varepsilon > 0$ and $\varepsilon \le F_0$, the nontrivial case has $F_0 > 0$, and convexity implies $\nabla f(x^0) \ne 0$. At the first inner iteration of phase zero, $x_0 = v_0 = y_0 = x^0$. Convexity and Cauchy–Schwarz give

$$F_0 \le \langle \nabla f(x^0), x^0 - x^* \rangle \le \left\|\nabla f(x^0)\right\| \widetilde{R}. \tag{30}$$

The descent inequality equation 12 for this first step, together with $f(x_1) \ge f^*$ and $M_0 \le \tau C_\star(H_0 + H_1 F_0)$, gives

$$\left\|\nabla f(x^0)\right\|^2 \le 2M_0 F_0 \le 2\tau C_\star(H_0 + H_1 F_0)F_0. \tag{31}$$

Squaring equation 30 and substituting equation 31 yields

$$F_0^2 \le 2\tau C_\star \widetilde{R}^2(H_0 + H_1 F_0)F_0, \tag{32}$$

and division of equation 32 by $F_0^2 > 0$ shows that

$$\widetilde{R}^2\left(\frac{H_0}{F_0} + H_1\right) \ge \frac{1}{2\tau C_\star}. \tag{33}$$

Consequently, equation 29 and equation 33 give $T_0 \ge \sqrt{2/\theta} \ge \sqrt{2}$. Since $\Delta_s = F_0/2^s$ is nonincreasing, $T_s \ge T_0 \ge \sqrt{2}$ for every phase. It follows that $N_s = \lceil T_s \rceil \le T_s + 1 \le 2T_s$. Therefore

$$\begin{aligned} N_{\mathrm{tot}} &\le 4\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}} \sum_{s=0}^{S-1} \sqrt{\frac{H_0}{\Delta_s} + H_1} \\ &\le 4\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}} \left(\sqrt{H_0}\sum_{s=0}^{S-1}\frac{1}{\sqrt{\Delta_s}} + S\sqrt{H_1}\right), \end{aligned} \tag{34}$$

where the second inequality in equation 34 uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$. The first sum is geometric:

$$\begin{aligned} \sum_{s=0}^{S-1}\frac{1}{\sqrt{\Delta_s}} &= \frac{1}{\sqrt{F_0}}\sum_{s=0}^{S-1} 2^{s/2} \\ &= \frac{2^{S/2} - 1}{(\sqrt{2} - 1)\sqrt{F_0}} \\ &\le \frac{\sqrt{2}}{\sqrt{2} - 1}\frac{1}{\sqrt{\varepsilon}}. \end{aligned} \tag{35}$$

The last inequality in equation 35 follows from $S = \lceil \log_2(F_0/\varepsilon) \rceil$, which implies $2^S \le 2F_0/\varepsilon$ whenever $S > 0$; for $S = 0$, the sum is empty. Finally, $S \le \log_2(2F_0/\varepsilon)$. Thus, with the standard convention that a logarithmic complexity factor is truncated below by one, $S = O(\log(F_0/\varepsilon))$. Substituting equation 35 and $S = O(\log(F_0/\varepsilon))$ into equation 34 gives

$$N_{\mathrm{tot}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta\varepsilon}} + \widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}}\log\frac{F_0}{\varepsilon}\right).$$

□

D Proof of Theorem 6

Theorem 6 (Coordinate complexity). *Suppose Assumptions 1 and 2 hold, and let $\varepsilon \in (0, F_0]$. Define $\mathcal{N}(\Delta) := \left\lceil 2d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta} + H_1}\right\rceil$. Then Algorithm 1, with Algorithm 3 as its inner method, returns a point z_S satisfying $\mathbb{E}[f(z_S) - f(x^*)] \le \varepsilon$. Moreover, the total number of iterations satisfies*

$$N_{\mathrm{tot}}^{\mathrm{coord}} = O\left(d\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta\varepsilon}} + d\widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}}\log\frac{F_0}{\varepsilon}\right).$$

Proof. The proof has five main steps. First, Lemma 10 guarantees that Algorithm 3 satisfies $|\nabla_i f(y_k)|/M_k \le r_1$ for every coordinate i, so the local model yields equation 40. Second, exact segment minimization and the conditional moments of uniform sampling give the coupling and potential estimates equation 46–equation 51. Third, growth of the accelerated weights turns the potential estimate into the one-phase guarantee equation 59. Fourth, the phase-length rule in Theorem 6 gives the expected contraction equation 65. Finally, we iterate this contraction and sum the phase lengths.

Coordinate descent. Consider one run of Algorithm 3, initialized at $x_0 = v_0 = z$ and $A_0 = 0$. Throughout the analysis of this run, z is fixed and expectations are taken over the coordinates sampled within the run. We first verify that every coordinate step lies inside the radius of the local quadratic model. Recall that $r_1 := 1/((2\sqrt{3} + \sqrt{6})\sqrt{H_1})$ when $H_1 > 0$, while $r_1 := \infty$ when $H_1 = 0$. When $H_1 > 0$, Lemma 10, the inequality $|\nabla_i f(y_k)| \le \|\nabla f(y_k)\|$, and $M_k = \tau C_\star(H_0 + H_1 F(y_k))$ give, for every coordinate i,

$$\frac{|\nabla_i f(y_k)|}{M_k} \le \frac{\|\nabla f(y_k)\|}{\tau C_\star(H_0 + H_1 F(y_k))} \le \frac{r_1}{\tau} \le r_1. \tag{36}$$

The first inequality in equation 36 uses $|\nabla_i f(y_k)| \le \|\nabla f(y_k)\|$. For the second inequality in equation 36, Lemma 10 gives $\|\nabla f(y_k)\|/r_1 \le C_\star(H_0 + H_1 F(y_k))$. Multiplying by r_1 and dividing by $\tau C_\star(H_0 + H_1 F(y_k))$ yields the factor r_1/τ. The last inequality in equation 36 follows from $\tau \ge 1$. When $H_1 = 0$, the conclusion is automatic because $r_1 = \infty$. Moreover, $\tau C_\star \ge 2$ implies $M_k \ge 2(H_0 + H_1 F(y_k))$.

After coordinate i_k is sampled, the primal update satisfies

$$\begin{aligned}\|x_{k+1} - y_k\| &= \left\| -\frac{\nabla_{i_k} f(y_k)}{M_k} e_{i_k} \right\| \\ &= \frac{|\nabla_{i_k} f(y_k)|}{M_k} \le r_1.\end{aligned} \tag{37}$$

The first equality in equation 37 follows by subtracting y_k from the update defining x_{k+1}; the second uses the absolute homogeneity of the norm and $\|e_{i_k}\| = 1$; and the last inequality follows from equation 36 with $i = i_k$. Hence the local model is valid with base point y_k and trial point x_{k+1}. It gives

$$\begin{aligned}f(x_{k+1}) &\le f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \big(H_0 + H_1 F(y_k)\big)\|x_{k+1} - y_k\|^2 \\ &= f(y_k) - \frac{(\nabla_{i_k} f(y_k))^2}{M_k} + \frac{H_0 + H_1 F(y_k)}{M_k^2}(\nabla_{i_k} f(y_k))^2.\end{aligned} \tag{38}$$

The equality follows by substituting $x_{k+1} - y_k = -M_k^{-1}\nabla_{i_k} f(y_k) e_{i_k}$, which gives

$$\langle \nabla f(y_k), x_{k+1} - y_k \rangle = -\frac{(\nabla_{i_k} f(y_k))^2}{M_k}, \qquad \|x_{k+1} - y_k\|^2 = \frac{(\nabla_{i_k} f(y_k))^2}{M_k^2}.$$

Since $\tau C_\star \geq 2$, the definition of M_k implies $M_k \geq 2(H_0 + H_1 F(y_k))$, and therefore

$$\frac{H_0 + H_1 F(y_k)}{M_k^2} \leq \frac{1}{2M_k}. \tag{39}$$

Indeed, dividing $H_0 + H_1 F(y_k) \leq M_k/2$ by $M_k^2 > 0$ gives equation 39. Substituting equation 39 into equation 38 and subtracting f^* from both sides gives the coordinate descent inequality

$$F(x_{k+1}) \leq F(y_k) - \frac{(\nabla_{i_k} f(y_k))^2}{2M_k}. \tag{40}$$

The choice $\beta = 1$ is feasible in the segment-relaxation problem and gives $v_k + \beta(x_k - v_k) = x_k$. Hence $f(y_k) = \min_{\beta \in [0,1]} f(v_k + \beta(x_k - v_k)) \leq f(x_k)$. Combining this observation with equation 40 shows, for every realization of i_k, that

$$f(x_{k+1}) \leq f(y_k) \leq f(x_k). \tag{41}$$

Thus both the primal iterates and the interpolation points remain below the objective value at the beginning of the inner run.

Conditional accelerated potential. We next establish the accelerated coupling. Let $\mathcal{F}_k$ denote the σ-algebra generated by the coordinate choices made before iteration k. The points x_k, v_k, y_k, the curvature M_k, and the weights a_{k+1}, A_{k+1} are determined before i_k is sampled and are therefore $\mathcal{F}_k$-measurable. Recall that

$$y_k \in \operatorname*{arg\,min}_{\beta \in [0,1]} f\big(v_k + \beta(x_k - v_k)\big),$$

define

$$w_k := \frac{A_k x_k + a_{k+1} v_k}{A_{k+1}}.$$

Because $A_{k+1} = A_k + a_{k+1}$, the coefficients in w_k are nonnegative and sum to one. Thus w_k lies on the segment minimized in the definition of y_k, and the first-order optimality condition for this convex one-dimensional problem implies

$$\langle \nabla f(y_k), w_k - y_k \rangle \geq 0. \tag{42}$$

Convexity of f, applied at y_k with comparison points x_k and x^*, respectively, gives

$$\begin{aligned} f(y_k) - f(x_k) &\leq \langle \nabla f(y_k), y_k - x_k \rangle, \\ F(y_k) &\leq \langle \nabla f(y_k), y_k - x^* \rangle. \end{aligned} \tag{43}$$

The first inequality in equation 43 is obtained by rearranging $f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle$. Similarly, the second inequality in equation 43 follows by rearranging $f^* = f(x^*) \geq f(y_k) + \langle \nabla f(y_k), x^* - y_k \rangle$. Because $A_k \geq 0$ and $a_{k+1} > 0$, multiplying the two inequalities in equation 43 by A_k and a_{k+1}, respectively, preserves their directions. Adding them gives

$$\begin{aligned} A_{k+1} F(y_k) - A_k F(x_k) &= A_k\big(f(y_k) - f(x_k)\big) + a_{k+1} F(y_k) \\ &\leq \langle \nabla f(y_k), A_k(y_k - x_k) + a_{k+1}(y_k - x^*) \rangle. \end{aligned} \tag{44}$$

To verify the equality in equation 44, use $A_{k+1} = A_k + a_{k+1}$ and $F(y_k) - F(x_k) = f(y_k) - f(x_k)$:

$$\begin{aligned} A_{k+1} F(y_k) - A_k F(x_k) &= (A_k + a_{k+1}) F(y_k) - A_k F(x_k) \\ &= A_k\big(F(y_k) - F(x_k)\big) + a_{k+1} F(y_k) \\ &= A_k\big(f(y_k) - f(x_k)\big) + a_{k+1} F(y_k). \end{aligned}$$

The inequality then follows by substituting the two convexity bounds and using the linearity of the inner product to combine their right-hand sides.

The vector inside the last inner product in equation 44 can be written as

$$\begin{aligned} A_k(y_k - x_k) + a_{k+1}(y_k - x^*) = a_{k+1}(v_k - x^*) \\ + A_{k+1}(y_k - w_k). \end{aligned} \tag{45}$$

Indeed, substituting the definition of w_k and expanding the right-hand side of equation 45 gives

$$\begin{aligned} &a_{k+1}(v_k - x^*) + A_{k+1}\left(y_k - \frac{A_k x_k + a_{k+1} v_k}{A_{k+1}}\right) \\ &= a_{k+1}v_k - a_{k+1}x^* + A_{k+1}y_k - A_k x_k - a_{k+1}v_k \\ &= A_{k+1}y_k - A_k x_k - a_{k+1}x^* \\ &= A_k(y_k - x_k) + a_{k+1}(y_k - x^*), \end{aligned}$$

where the last equality again uses $A_{k+1} = A_k + a_{k+1}$. Substituting equation 45 into equation 44 gives

$$\begin{aligned} A_{k+1}F(y_k) - A_k F(x_k) &\le a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle + A_{k+1}\left\langle \nabla f(y_k), y_k - w_k \right\rangle \\ &\le a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle. \end{aligned} \tag{46}$$

The second inequality in equation 46 follows from equation 42: that condition is $\langle \nabla f(y_k), w_k - y_k \rangle \ge 0$, or equivalently $\langle \nabla f(y_k), y_k - w_k \rangle \le 0$. Multiplication by $A_{k+1} > 0$ preserves this inequality, so the second term in the first line of equation 46 is nonpositive.

Uniform sampling and $\mathcal{F}_k$-measurability of y_k imply

$$\begin{aligned} \mathbb{E}[d\nabla_{i_k} f(y_k) e_{i_k} \,|\, \mathcal{F}_k] &= \sum_{i=1}^{d} \frac{1}{d}\, d\nabla_i f(y_k) e_i = \nabla f(y_k), \\ \mathbb{E}\left[(\nabla_{i_k} f(y_k))^2 \,\middle|\, \mathcal{F}_k\right] &= \sum_{i=1}^{d} \frac{1}{d}(\nabla_i f(y_k))^2 = \frac{1}{d}\left\|\nabla f(y_k)\right\|^2. \end{aligned} \tag{47}$$

Conditionally on $\mathcal{F}_k$, each coordinate is selected with probability $\Pr(i_k = i \mid \mathcal{F}_k) = 1/d$. In the first identity of equation 47, scaling the sampled vector by d compensates for the sampling probability $1/d$, since $(1/d)d = 1$. Therefore $\sum_{i=1}^{d} \nabla_i f(y_k) e_i = \nabla f(y_k)$, so $d\nabla_{i_k} f(y_k) e_{i_k}$ is an unbiased estimator of the full gradient. The squared partial derivative is not multiplied by d; consequently, its uniform average retains the factor $1/d$: $\mathbb{E}[(\nabla_{i_k} f(y_k))^2 \mid \mathcal{F}_k] = (1/d)\sum_{i=1}^{d}(\nabla_i f(y_k))^2 = (1/d)\left\|\nabla f(y_k)\right\|^2$. Multiplying equation 40 by the $\mathcal{F}_k$-measurable nonnegative weight A_{k+1} and taking conditional expectation, we may move A_{k+1}, M_k, and $F(y_k)$ outside the expectation. Using the second identity in equation 47 gives

$$\begin{aligned} \mathbb{E}[A_{k+1}F(x_{k+1}) \,|\, \mathcal{F}_k] &\le A_{k+1}F(y_k) - \frac{A_{k+1}}{2M_k}\mathbb{E}\left[(\nabla_{i_k} f(y_k))^2 \,\middle|\, \mathcal{F}_k\right] \\ &= A_{k+1}F(y_k) - \frac{A_{k+1}}{2dM_k}\left\|\nabla f(y_k)\right\|^2. \end{aligned} \tag{48}$$

Expanding the squared norm in

$$v_{k+1} = v_k - da_{k+1}\nabla_{i_k} f(y_k) e_{i_k}$$

gives, before taking expectation,

$$\begin{aligned} \frac{1}{2}\left\|v_{k+1} - x^*\right\|^2 = \frac{1}{2}\left\|v_k - x^*\right\|^2 - da_{k+1}\nabla_{i_k} f(y_k)\left\langle e_{i_k}, v_k - x^* \right\rangle \\ + \frac{d^2 a_{k+1}^2}{2}(\nabla_{i_k} f(y_k))^2. \end{aligned}$$

Conditioning on $\mathcal{F}_k$, the first identity in equation 47 turns the middle term into $-a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle$, while the second identity in equation 47 turns the last term into $(da_{k+1}^2/2)\left\|\nabla f(y_k)\right\|^2$. Hence

$$\begin{aligned} \mathbb{E}\left[\frac{1}{2}\left\|v_{k+1} - x^*\right\|^2 \,\middle|\, \mathcal{F}_k\right] = \frac{1}{2}\left\|v_k - x^*\right\|^2 - a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle \\ + \frac{da_{k+1}^2}{2}\left\|\nabla f(y_k)\right\|^2. \end{aligned} \tag{49}$$

Define the nonnegative potential

$$\Phi_k := A_k F(x_k) + \frac{1}{2}\|v_k - x^*\|^2 .$$

It is nonnegative because $A_k \geq 0$, $F(x_k) \geq 0$, and the squared norm is nonnegative. Adding equation 48 and equation 49, and then applying equation 46 gives

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \leq \Phi_k + \left(\frac{da_{k+1}^2}{2} - \frac{A_{k+1}}{2dM_k}\right)\|\nabla f(y_k)\|^2 .$$

Indeed, rearranging equation 46 yields

$$A_{k+1}F(y_k) - a_{k+1}\langle \nabla f(y_k), v_k - x^*\rangle \leq A_k F(x_k).$$

Thus equation 46 bounds the two linear terms obtained after adding equation 48 and equation 49 by $A_k F(x_k)$; together with $\frac{1}{2}\|v_k - x^*\|^2$, they form Φ_k. By the weight equation in Algorithm 3, $d^2 M_k a_{k+1}^2 = \theta A_{k+1}$. Hence

$$\begin{aligned}\frac{da_{k+1}^2}{2} - \frac{A_{k+1}}{2dM_k} &= \frac{\theta A_{k+1}}{2dM_k} - \frac{A_{k+1}}{2dM_k} \\ &= -\frac{(1-\theta)A_{k+1}}{2dM_k} \leq 0,\end{aligned} \tag{50}$$

where the last inequality in equation 50 uses $\theta \in (0, 1]$. It follows that $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \leq \Phi_k$. Taking total expectation and applying the tower property successively for $k = 0, \ldots, N-1$ gives

$$\mathbb{E}[\Phi_N \mid z] \leq \Phi_0 = \frac{1}{2}\|z - x^*\|^2 , \tag{51}$$

because $A_0 = 0$ and $v_0 = z$. Indeed, at each iteration the tower property gives

$$\mathbb{E}[\Phi_{k+1} \mid z] = \mathbb{E}[\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \mid z] \leq \mathbb{E}[\Phi_k \mid z]. \tag{52}$$

Iterating equation 52 yields equation 51.

One-phase estimate. We now lower bound the accelerated weight A_N along every realization. Since $A_{k+1} - A_k = a_{k+1} > 0$, we have $A_k \leq A_{k+1}$, and therefore

$$\begin{aligned}\sqrt{A_{k+1}} - \sqrt{A_k} &= \frac{a_{k+1}}{\sqrt{A_{k+1}} + \sqrt{A_k}} \\ &\geq \frac{a_{k+1}}{2\sqrt{A_{k+1}}} = \frac{\sqrt{\theta}}{2d\sqrt{M_k}}.\end{aligned} \tag{53}$$

The first equality in equation 53 is the difference-of-squares identity, the inequality uses $\sqrt{A_k} \leq \sqrt{A_{k+1}}$, and the final equality follows from $d^2 M_k a_{k+1}^2 = \theta A_{k+1}$. Summing equation 53 over $k = 0, \ldots, N-1$ and using $A_0 = 0$ yields

$$\sqrt{A_N} \geq \frac{\sqrt{\theta}}{2d}\sum_{k=0}^{N-1}\frac{1}{\sqrt{M_k}}. \tag{54}$$

The pathwise monotonicity in equation 41, together with $x_0 = z$, gives $F(y_k) \leq F(z)$. Since $M_k = \tau C_\star(H_0 + H_1 F(y_k))$ and $H_1 \geq 0$, it follows that

$$M_k \leq \tau C_\star\big(H_0 + H_1 F(z)\big). \tag{55}$$

Since both sides of equation 55 are positive and $t \mapsto t^{-1/2}$ is decreasing on $(0, \infty)$, taking reciprocal square roots in equation 55 gives

$$\frac{1}{\sqrt{M_k}} \geq \frac{1}{\sqrt{\tau C_\star(H_0 + H_1 F(z))}}. \tag{56}$$

Substituting equation 56 into every summand of equation 54 gives

$$\sqrt{A_N} \geq \frac{\sqrt{\theta}\, N}{2d\sqrt{\tau C_\star(H_0 + H_1 F(z))}}. \tag{57}$$

Both sides of equation 57 are nonnegative, so squaring preserves the inequality and gives

$$A_N \geq \frac{\theta N^2}{4d^2\tau C_\star\big(H_0 + H_1 F(z)\big)}. \tag{58}$$

Denote the deterministic right-hand side of equation 58 by $\underline{A}_N(z)$. Equation equation 58 holds for every realization of the sampled coordinates. Since $F(x_N) \geq 0$, multiplication preserves the inequality: $\underline{A}_N(z)F(x_N) \leq A_N F(x_N)$. Moreover, $\Phi_N = A_N F(x_N) + \frac{1}{2}\|v_N - x^*\|^2 \geq A_N F(x_N)$. Taking conditional expectation therefore gives

$$\begin{aligned}\frac{\theta N^2}{4d^2\tau C_\star(H_0 + H_1 F(z))}\mathbb{E}[F(x_N) \mid z] &\leq \mathbb{E}[A_N F(x_N) \mid z] \\ &\leq \mathbb{E}[\Phi_N \mid z] \leq \frac{1}{2}\|z - x^*\|^2.\end{aligned}$$

The last inequality follows from equation 51. Rearranging proves the phase estimate

$$\mathbb{E}[F(x_N) \mid z] \leq \frac{2d^2\tau C_\star \|z - x^*\|^2}{\theta N^2}\big(H_0 + H_1 F(z)\big). \tag{59}$$

Restart contraction. We apply equation 59 to the restart wrapper. Recall

$$\widetilde{R} = \sup_{f(x)\leq f(x^0)} \|x - x^*\| < \infty.$$

Its finiteness follows from Lemma 11. The pathwise monotonicity equation 41 implies $f(z_s) \leq f(x^0)$ for every realization, so z_s belongs to the initial sublevel set and $\|z_s - x^*\| \leq \widetilde{R}$. Applying equation 59 with $z = z_s$, $N = N_s$, and $x_N = z_{s+1}$, conditionally on the realized phase start z_s, gives

$$\mathbb{E}[F(z_{s+1}) \mid z_s] \leq \frac{2d^2\tau C_\star \widetilde{R}^2}{\theta N_s^2}\big(H_0 + H_1 F(z_s)\big). \tag{60}$$

Suppose inductively that $\mathbb{E}[F(z_s)] \leq \Delta_s$, where $\Delta_s = F_0/2^s$. Taking total expectation in equation 60 and using $\mathbb{E}[\mathbb{E}[F(z_{s+1}) \mid z_s]] = \mathbb{E}[F(z_{s+1})]$ gives

$$\begin{aligned}\mathbb{E}[F(z_{s+1})] &\leq \frac{2d^2\tau C_\star \widetilde{R}^2}{\theta N_s^2}\big(H_0 + H_1 \mathbb{E}[F(z_s)]\big) \\ &\leq \frac{2d^2\tau C_\star \widetilde{R}^2}{\theta N_s^2}\big(H_0 + H_1 \Delta_s\big).\end{aligned} \tag{61}$$

Here N_s and all other prefactors are deterministic. The first inequality in equation 61 uses the linearity of expectation, and the second uses $\mathbb{E}[F(z_s)] \leq \Delta_s$ together with $H_1 \geq 0$. The phase-length rule gives

$$N_s \geq 2d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta_s} + H_1}, \tag{62}$$

and hence

$$N_s^2 \geq \frac{4d^2\tau C_\star \widetilde{R}^2}{\theta}\frac{H_0 + H_1\Delta_s}{\Delta_s}. \tag{63}$$

Equation equation 63 follows by squaring equation 62 and using $H_0/\Delta_s + H_1 = (H_0 + H_1\Delta_s)/\Delta_s$. Since all factors are positive, equation 63 implies

$$\frac{1}{N_s^2} \leq \frac{\theta\Delta_s}{4d^2\tau C_\star \widetilde{R}^2(H_0 + H_1\Delta_s)}. \tag{64}$$

Substituting equation 64 into equation 61 yields

$$\begin{aligned}\mathbb{E}[F(z_{s+1})] &\leq \frac{2d^2\tau C_\star \widetilde{R}^2}{\theta}(H_0 + H_1\Delta_s)\frac{\theta\Delta_s}{4d^2\tau C_\star \widetilde{R}^2(H_0 + H_1\Delta_s)} \\ &= \frac{\Delta_s}{2} = \Delta_{s+1}.\end{aligned} \tag{65}$$

The base case holds because $z_0 = x^0$ and $F(z_0) = F_0 = \Delta_0$. Repeated application of equation 65 therefore proves $\mathbb{E}[F(z_s)] \leq \Delta_s$ for every phase. For $S = \lceil \log_2(F_0/\varepsilon) \rceil$, we have $\Delta_S = F_0/2^S \leq \varepsilon$, and consequently

$$\mathbb{E}[f(z_S) - f(x^*)] = \mathbb{E}[F(z_S)] \leq \Delta_S \leq \varepsilon.$$

Total complexity. It remains to sum the phase lengths. Denote the quantity inside the ceiling by

$$T_s := 2d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{\Delta_s}+H_1}, \qquad N_s = \lceil T_s \rceil. \tag{66}$$

We first verify that the ceiling changes the sum only by an absolute factor. In the nontrivial case $F_0 > 0$, convexity and Cauchy–Schwarz imply

$$F_0 \le \left\langle \nabla f(x^0), x^0 - x^* \right\rangle \le \left\|\nabla f(x^0)\right\| \, \widetilde{R}. \tag{67}$$

The first inequality in equation 67 is the convexity relation $f(x^*) \ge f(x^0) + \left\langle \nabla f(x^0), x^* - x^0 \right\rangle$; the second inequality in equation 67 is Cauchy–Schwarz together with $\left\|x^0 - x^*\right\| \le \widetilde{R}$. Lemma 10, applied at x^0, gives

$$\frac{\left\|\nabla f(x^0)\right\|}{\tau C_\star(H_0 + H_1 F_0)} \le \frac{r_1}{\tau} \le r_1. \tag{68}$$

By equation 68, the analytical full-gradient step

$$x^0 - \frac{\nabla f(x^0)}{\tau C_\star(H_0 + H_1 F_0)}$$

lies inside the local-model radius. Since $\tau C_\star(H_0 + H_1 F_0) \ge 2(H_0 + H_1 F_0)$, the same calculation as in equation 40, now using the full gradient, gives

$$f\left(x^0 - \frac{\nabla f(x^0)}{\tau C_\star(H_0 + H_1 F_0)}\right) \le f(x^0) - \frac{\left\|\nabla f(x^0)\right\|^2}{2\tau C_\star(H_0 + H_1 F_0)}. \tag{69}$$

The left-hand side of equation 69 is at least f^*. Subtracting f^* and rearranging therefore yields

$$\left\|\nabla f(x^0)\right\|^2 \le 2\tau C_\star(H_0 + H_1 F_0) F_0. \tag{70}$$

Squaring equation 67 and substituting equation 70 gives

$$F_0^2 \le \left\|\nabla f(x^0)\right\|^2 \widetilde{R}^2 \le 2\tau C_\star(H_0 + H_1 F_0) F_0 \widetilde{R}^2. \tag{71}$$

Dividing equation 71 by $F_0^2 > 0$ and rearranging gives

$$\widetilde{R}^2\left(\frac{H_0}{F_0} + H_1\right) \ge \frac{1}{2\tau C_\star}. \tag{72}$$

Since $\Delta_0 = F_0$, substituting equation 72 into equation 66 with $s = 0$ gives

$$\begin{aligned} T_0 &= 2d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0}{F_0}+H_1} \\ &\ge 2d\sqrt{\frac{\tau C_\star}{\theta}}\frac{1}{\sqrt{2\tau C_\star}} = d\sqrt{\frac{2}{\theta}} \ge \sqrt{2}, \end{aligned} \tag{73}$$

where the last inequality in equation 73 uses $d \ge 1$ and $\theta \le 1$. Since Δ_s is nonincreasing, $H_0/\Delta_s + H_1$ and hence T_s are nondecreasing, so $T_s \ge T_0 \ge 1$ for every phase. Consequently, $N_s = \lceil T_s \rceil \le T_s + 1 \le 2T_s$, and

$$\begin{aligned} N_{\mathrm{tot}}^{\mathrm{coord}} &= \sum_{s=0}^{S-1} N_s \\ &\le 4d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sum_{s=0}^{S-1}\sqrt{\frac{H_0}{\Delta_s}+H_1} \\ &\le 4d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\left(\sqrt{H_0}\sum_{s=0}^{S-1}\frac{1}{\sqrt{\Delta_s}} + S\sqrt{H_1}\right). \end{aligned} \tag{74}$$

The last inequality in equation 74 uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$. Since $\Delta_s = F_0/2^s$,

$$\begin{aligned}\sum_{s=0}^{S-1} \frac{1}{\sqrt{\Delta_s}} &= \frac{1}{\sqrt{F_0}} \sum_{s=0}^{S-1} 2^{s/2} \\ &= \frac{2^{S/2}-1}{(\sqrt{2}-1)\sqrt{F_0}} \\ &\le \frac{\sqrt{2}}{\sqrt{2}-1} \frac{1}{\sqrt{\varepsilon}}.\end{aligned} \tag{75}$$

The last inequality in equation 75 follows from $S = \lceil \log_2(F_0/\varepsilon) \rceil$, which implies $2^S \le 2F_0/\varepsilon$ when $S > 0$; if $S = 0$, the sum is empty. Moreover, $S \le \log_2(2F_0/\varepsilon)$, so, with the usual convention that a logarithmic complexity factor is truncated below by one, $S = O(\log(F_0/\varepsilon))$. Substituting equation 75 and $S = O(\log(F_0/\varepsilon))$ into equation 74 gives

$$N_{\mathrm{tot}}^{\mathrm{coord}} = O\left(d\widetilde{R}\sqrt{\frac{\tau C_\star H_0}{\theta\varepsilon}} + d\widetilde{R}\sqrt{\frac{\tau C_\star H_1}{\theta}} \log \frac{F_0}{\varepsilon} \right).$$

□

E Proof of Theorem 8

Theorem 8 (Non-uniform coordinate complexity). *Suppose Assumptions 1 and 7 hold, and let $\varepsilon \in (0, F_0]$. Define $\mathcal{N}(\Delta) := \left\lceil 2\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}(S_{1/2}^{(0)}/\sqrt{\Delta} + S_{1/2}^{(1)}) \right\rceil$. Then Algorithm 1, with Algorithm 4 as its inner method, returns a point z_S satisfying $\mathbb{E}[f(z_S) - f(x^*)] \le \varepsilon$. Moreover, the total number of coordinate-gradient calls, equivalently iterations, satisfies*

$$N_{\mathrm{tot}}^{\mathrm{nonunif}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}} \left(\frac{S_{1/2}^{(0)}}{\sqrt{\varepsilon}} + S_{1/2}^{(1)} \log \frac{F_0}{\varepsilon} \right) \right).$$

Proof. The proof has five main steps. First, the coordinate statement of Lemma 10 shows that every update of Algorithm 4 lies inside the corresponding radius $r_{1,i}$. This justifies the local quadratic model and gives the coordinate descent and monotonicity estimates equation 80 and equation 81. Second, exact segment minimization, convexity, and the conditional moments of the importance-sampled coordinate yield the coupling inequality equation 86 and the potential estimate equation 92. Third, we control the probability normalizer $\mathcal{M}_k$, lower bound the accelerated weight A_N, and derive the one-phase estimate equation 97. Fourth, Jensen's inequality and the phase-length rule imply the expected contraction equation 104. Finally, induction over the restart phases gives the desired accuracy, and summing their lengths gives the stated complexity.

Coordinate descent. Consider one run of Algorithm 4, initialized at $x_0 = v_0 = z$ and $A_0 = 0$. Throughout the analysis of this run, the phase start z is fixed and expectations are taken over the sampled coordinates. For every coordinate i, Algorithm 4 sets $M_{k,i} = \tau C_\star(H_{0,i} + H_{1,i}F(y_k))$. The coordinate statement of Lemma 10 therefore gives

$$\begin{aligned}\frac{|\nabla_i f(y_k)|}{M_{k,i}} &= \frac{|\nabla_i f(y_k)|}{\tau C_\star(H_{0,i} + H_{1,i}F(y_k))} \\ &\le \frac{r_{1,i}}{\tau} \le r_{1,i}.\end{aligned} \tag{76}$$

Indeed, Lemma 10 states that $|\nabla_i f(y_k)|/r_{1,i} \le C_\star(H_{0,i} + H_{1,i}F(y_k))$. Multiplying this inequality by $r_{1,i}$ and dividing by the positive denominator in the first line of equation 76 gives the factor $r_{1,i}/\tau$; the last inequality uses $\tau \ge 1$. When $H_{1,i} = 0$, the same conclusion is automatic under the convention $r_{1,i} = \infty$. If $F(y_k) = 0$, then y_k is already optimal and the inner run may terminate. On

every nonterminal iteration $F(y_k) > 0$; since $H_{1,i} > 0$, all $M_{k,i}$, $\mathcal{M}_k$, and $p_{k,i}$ are then positive. The other part of Lemma 10, together with $\tau \geq 1$, also gives

$$M_{k,i} \geq 2\big(H_{0,i} + H_{1,i}F(y_k)\big). \tag{77}$$

After coordinate i_k is sampled, subtracting y_k from the primal update gives

$$\begin{aligned} \|x_{k+1} - y_k\| &= \left\| -\frac{\nabla_{i_k} f(y_k)}{M_{k,i_k}} e_{i_k} \right\| \\ &= \frac{|\nabla_{i_k} f(y_k)|}{M_{k,i_k}} \leq r_{1,i_k}. \end{aligned} \tag{78}$$

The second equality in equation 78 uses $\|e_{i_k}\| = 1$, and its last inequality is equation 76 with $i = i_k$. Hence the coordinate local model in Assumption 7 is valid with base point y_k and scalar displacement $-\nabla_{i_k} f(y_k)/M_{k,i_k}$. Substituting this displacement into that model yields

$$f(x_{k+1}) \leq f(y_k) - \frac{(\nabla_{i_k} f(y_k))^2}{M_{k,i_k}} + \frac{H_{0,i_k} + H_{1,i_k}F(y_k)}{M_{k,i_k}^2}(\nabla_{i_k} f(y_k))^2. \tag{79}$$

The linear term in the local model equals $-M_{k,i_k}^{-1}(\nabla_{i_k} f(y_k))^2$, while the squared displacement equals $M_{k,i_k}^{-2}(\nabla_{i_k} f(y_k))^2$, which explains the two terms on the right-hand side of equation 79. By equation 77,

$$\frac{H_{0,i_k} + H_{1,i_k}F(y_k)}{M_{k,i_k}^2} \leq \frac{1}{2M_{k,i_k}}.$$

Substituting this bound into equation 79 and subtracting f^* from both sides gives

$$F(x_{k+1}) \leq F(y_k) - \frac{(\nabla_{i_k} f(y_k))^2}{2M_{k,i_k}}. \tag{80}$$

The segment defining y_k contains x_k, because $\beta = 1$ is feasible and gives $v_k + \beta(x_k - v_k) = x_k$. Consequently, $f(y_k) \leq f(x_k)$. Equation equation 80 also gives $f(x_{k+1}) \leq f(y_k)$, and therefore, for every realization of i_k,

$$f(x_{k+1}) \leq f(y_k) \leq f(x_k). \tag{81}$$

Importance-sampled potential. Let $\mathcal{F}_k$ be the σ-algebra generated by the coordinates sampled before iteration k. The points x_k, v_k, y_k, all curvatures $M_{k,i}$, the probabilities $p_{k,i}$, and the weights a_{k+1}, A_{k+1} are determined before i_k is sampled and are therefore $\mathcal{F}_k$-measurable. Define

$$w_k := \frac{A_k x_k + a_{k+1} v_k}{A_{k+1}}.$$

Since $A_{k+1} = A_k + a_{k+1}$, the coefficients defining w_k are nonnegative and sum to one. Thus w_k lies on the segment minimized in the definition of y_k. The first-order optimality condition for this convex one-dimensional problem gives

$$\langle \nabla f(y_k), w_k - y_k \rangle \geq 0. \tag{82}$$

Convexity of f, applied at y_k with comparison points x_k and x^*, gives

$$\begin{aligned} f(y_k) - f(x_k) &\leq \langle \nabla f(y_k), y_k - x_k \rangle, \\ F(y_k) &\leq \langle \nabla f(y_k), y_k - x^* \rangle. \end{aligned} \tag{83}$$

The first inequality is obtained by rearranging the convexity inequality at x_k, and the second by rearranging the same inequality at the minimizer x^*. Multiplying the two inequalities in equation 83 by A_k and a_{k+1}, respectively, and adding them yields

$$\begin{aligned} A_{k+1}F(y_k) - A_k F(x_k) &= A_k\big(f(y_k) - f(x_k)\big) + a_{k+1}F(y_k) \\ &\leq \langle \nabla f(y_k), A_k(y_k - x_k) + a_{k+1}(y_k - x^*) \rangle. \end{aligned} \tag{84}$$

The equality in equation 84 follows from $A_{k+1} = A_k + a_{k+1}$ and $F(y_k) - F(x_k) = f(y_k) - f(x_k)$. The vector in its final inner product satisfies

$$A_k(y_k - x_k) + a_{k+1}(y_k - x^*) = a_{k+1}(v_k - x^*) + A_{k+1}(y_k - w_k). \tag{85}$$

Indeed, substituting the definition of w_k into the right-hand side of equation 85 gives

$$\begin{aligned} &a_{k+1}(v_k - x^*) + A_{k+1}y_k - A_k x_k - a_{k+1}v_k \\ &\quad = A_{k+1}y_k - A_k x_k - a_{k+1}x^* = A_k(y_k - x_k) + a_{k+1}(y_k - x^*). \end{aligned}$$

Substituting equation 85 into equation 84, and then using equation 82, gives the coupling inequality

$$\begin{aligned} A_{k+1}F(y_k) - A_k F(x_k) &\le a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle + A_{k+1}\left\langle \nabla f(y_k), y_k - w_k \right\rangle \\ &\le a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle. \end{aligned} \tag{86}$$

The second inequality follows because equation 82 is equivalent to $\langle \nabla f(y_k), y_k - w_k \rangle \le 0$, and $A_{k+1} > 0$.

Conditionally on $\mathcal{F}_k$, coordinate i is sampled with probability $p_{k,i}$. Therefore,

$$\begin{aligned} \mathbb{E}\Big[p_{k,i_k}^{-1} \nabla_{i_k} f(y_k) e_{i_k} \,\Big|\, \mathcal{F}_k \Big] &= \sum_{i=1}^{d} p_{k,i} p_{k,i}^{-1} \nabla_i f(y_k) e_i = \nabla f(y_k), \\ \mathbb{E}\left[\frac{(\nabla_{i_k} f(y_k))^2}{M_{k,i_k}} \,\middle|\, \mathcal{F}_k \right] &= \sum_{i=1}^{d} \frac{p_{k,i}}{M_{k,i}} (\nabla_i f(y_k))^2, \\ \mathbb{E}\Big[p_{k,i_k}^{-2} (\nabla_{i_k} f(y_k))^2 \,\Big|\, \mathcal{F}_k \Big] &= \sum_{i=1}^{d} \frac{(\nabla_i f(y_k))^2}{p_{k,i}}. \end{aligned} \tag{87}$$

In the first identity, the factor $p_{k,i}^{-1}$ compensates for the sampling probability $p_{k,i}$, so the sampled vector is an unbiased estimator of the full gradient. In the last identity, one factor $p_{k,i}$ from the expectation combines with $p_{k,i}^{-2}$, leaving $p_{k,i}^{-1}$.

Multiplying equation 80 by the nonnegative, $\mathcal{F}_k$-measurable weight A_{k+1}, taking conditional expectation, and using the second identity in equation 87 gives

$$\begin{aligned} \mathbb{E}[A_{k+1}F(x_{k+1}) \,|\, \mathcal{F}_k] \le A_{k+1}F(y_k) \\ - \frac{A_{k+1}}{2} \sum_{i=1}^{d} \frac{p_{k,i}}{M_{k,i}} (\nabla_i f(y_k))^2. \end{aligned} \tag{88}$$

Expanding the squared distance after the auxiliary update gives, before taking expectation,

$$\begin{aligned} \frac{1}{2}\left\| v_{k+1} - x^* \right\|^2 = \frac{1}{2}\left\| v_k - x^* \right\|^2 - a_{k+1} p_{k,i_k}^{-1} \nabla_{i_k} f(y_k) \left\langle e_{i_k}, v_k - x^* \right\rangle \\ + \frac{a_{k+1}^2}{2} p_{k,i_k}^{-2} (\nabla_{i_k} f(y_k))^2. \end{aligned}$$

The first and third identities in equation 87 therefore imply

$$\begin{aligned} \mathbb{E}\left[\frac{1}{2}\left\| v_{k+1} - x^* \right\|^2 \,\middle|\, \mathcal{F}_k \right] = \frac{1}{2}\left\| v_k - x^* \right\|^2 - a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle \\ + \frac{a_{k+1}^2}{2} \sum_{i=1}^{d} \frac{(\nabla_i f(y_k))^2}{p_{k,i}}. \end{aligned} \tag{89}$$

Define the nonnegative potential

$$\Phi_k := A_k F(x_k) + \frac{1}{2}\left\| v_k - x^* \right\|^2.$$

Adding equation 88 and equation 89, and then rearranging equation 86 as

$$A_{k+1}F(y_k) - a_{k+1}\left\langle \nabla f(y_k), v_k - x^* \right\rangle \le A_k F(x_k),$$

gives

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k + \frac{1}{2} \sum_{i=1}^{d} \left(\frac{a_{k+1}^2}{p_{k,i}} - \frac{A_{k+1} p_{k,i}}{M_{k,i}} \right) (\nabla_i f(y_k))^2. \tag{90}$$

Recall that $\mathcal{M}_k := \sum_{j=1}^d \sqrt{M_{k,j}}$, so $p_{k,i} = \sqrt{M_{k,i}}/\mathcal{M}_k$. Together with $\mathcal{M}_k^2 a_{k+1}^2 = \theta A_{k+1}$, this implies that every coefficient in equation 90 satisfies

$$\begin{aligned} \frac{a_{k+1}^2}{p_{k,i}} - \frac{A_{k+1} p_{k,i}}{M_{k,i}} &= \frac{\mathcal{M}_k^2 a_{k+1}^2 - A_{k+1}}{\mathcal{M}_k \sqrt{M_{k,i}}} \\ &= -\frac{(1-\theta) A_{k+1}}{\mathcal{M}_k \sqrt{M_{k,i}}} \le 0. \end{aligned} \tag{91}$$

The first equality follows by substituting $p_{k,i} = \sqrt{M_{k,i}}/\mathcal{M}_k$ and placing the two fractions over the common denominator $\mathcal{M}_k \sqrt{M_{k,i}}$; the second uses the accelerated weight equation, and the final inequality uses $\theta \in (0, 1]$. Substituting equation 91 into equation 90 yields $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k$. The tower property then gives, for each k,

$$\mathbb{E}[\Phi_{k+1} \mid z] = \mathbb{E}[\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \mid z] \le \mathbb{E}[\Phi_k \mid z].$$

Applying this inequality successively and using $A_0 = 0$ and $v_0 = z$ gives

$$\mathbb{E}[\Phi_N \mid z] \le \Phi_0 = \frac{1}{2} \|z - x^*\|^2. \tag{92}$$

One-phase estimate. We next lower bound A_N along every realization. Since $A_{k+1} - A_k = a_{k+1} > 0$, we have $A_k \le A_{k+1}$, and hence

$$\begin{aligned} \sqrt{A_{k+1}} - \sqrt{A_k} &= \frac{a_{k+1}}{\sqrt{A_{k+1}} + \sqrt{A_k}} \\ &\ge \frac{a_{k+1}}{2\sqrt{A_{k+1}}} = \frac{\sqrt{\theta}}{2\mathcal{M}_k}. \end{aligned} \tag{93}$$

The first equality in equation 93 is the difference-of-squares identity, the inequality uses $\sqrt{A_k} \le \sqrt{A_{k+1}}$, and the final equality follows from $\mathcal{M}_k^2 a_{k+1}^2 = \theta A_{k+1}$.

The pathwise monotonicity in equation 81, together with $x_0 = z$, implies $F(y_k) \le F(z)$. Recall that $S_{1/2}^{(0)} := \sum_{i=1}^d \sqrt{H_{0,i}}$ and $S_{1/2}^{(1)} := \sum_{i=1}^d \sqrt{H_{1,i}}$; unlike $\mathcal{M}_k$, both quantities are fixed across iterations. Therefore

$$\begin{aligned} \mathcal{M}_k &= \sqrt{\tau C_\star} \sum_{i=1}^d \sqrt{H_{0,i} + H_{1,i} F(y_k)} \\ &\le \sqrt{\tau C_\star} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z)} \right). \end{aligned} \tag{94}$$

The inequality in equation 94 uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for every coordinate and then uses $F(y_k) \le F(z)$. Since all quantities are positive, taking reciprocals in equation 94 reverses the inequality. Substituting the resulting lower bound on $1/\mathcal{M}_k$ into equation 93, summing over $k = 0, \dots, N-1$, and using $A_0 = 0$, gives

$$\sqrt{A_N} \ge \frac{\sqrt{\theta}\, N}{2\sqrt{\tau C_\star} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z)} \right)}. \tag{95}$$

Both sides of equation 95 are nonnegative, so squaring preserves the inequality and yields

$$A_N \ge \frac{\theta N^2}{4\tau C_\star \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z)} \right)^2}. \tag{96}$$

The right-hand side of equation 96 depends on the fixed phase start z, but not on the coordinates sampled during the phase. Since $F(x_N) \ge 0$, multiplying equation 96 by $F(x_N)$ preserves the inequality. Moreover, $\Phi_N \ge A_N F(x_N)$. Taking conditional expectation and applying equation 92 therefore gives

$$\begin{aligned} &\frac{\theta N^2}{4\tau C_\star \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z)} \right)^2} \mathbb{E}[F(x_N) \mid z] \\ &\qquad \le \mathbb{E}[A_N F(x_N) \mid z] \le \mathbb{E}[\Phi_N \mid z] \le \frac{1}{2} \|z - x^*\|^2. \end{aligned}$$

Rearranging this chain gives the one-phase estimate

$$\mathbb{E}[F(x_N) \mid z] \le \frac{2\tau C_\star \|z - x^*\|^2}{\theta N^2} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z)}\right)^2. \tag{97}$$

Restart contraction. We now apply equation 97 to the restart wrapper. By equation 81, every restart point satisfies $f(z_s) \le f(x^0)$ for every realization. Hence Lemma 11 and equation 23 imply

$$\|z_s - x^*\| \le \widetilde{R}. \tag{98}$$

Applying equation 97 with $z = z_s$, $N = N_s$, and $x_N = z_{s+1}$, conditionally on z_s, and then using equation 98, gives

$$\mathbb{E}[F(z_{s+1}) \mid z_s] \le \frac{2\tau C_\star \widetilde{R}^2}{\theta N_s^2} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z_s)}\right)^2. \tag{99}$$

Suppose inductively that $\mathbb{E}F(z_s) \le \Delta_s$. Expanding the square, using Jensen's inequality $\mathbb{E}\sqrt{F(z_s)} \le \sqrt{\mathbb{E}F(z_s)}$, and then applying the induction hypothesis gives

$$\begin{aligned} &\mathbb{E}\left[\left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F(z_s)}\right)^2\right] \\ &\quad = \left(S_{1/2}^{(0)}\right)^2 + 2S_{1/2}^{(0)} S_{1/2}^{(1)} \mathbb{E}\sqrt{F(z_s)} + \left(S_{1/2}^{(1)}\right)^2 \mathbb{E}F(z_s) \\ &\quad \le \left(S_{1/2}^{(0)}\right)^2 + 2S_{1/2}^{(0)} S_{1/2}^{(1)} \sqrt{\Delta_s} + \left(S_{1/2}^{(1)}\right)^2 \Delta_s \\ &\quad = \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{\Delta_s}\right)^2. \end{aligned} \tag{100}$$

Taking total expectation in equation 99, using the tower property, and then substituting equation 100 yields

$$\mathbb{E}F(z_{s+1}) \le \frac{2\tau C_\star \widetilde{R}^2}{\theta N_s^2} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{\Delta_s}\right)^2. \tag{101}$$

By the definition of $N_s = \mathcal{N}(\Delta_s)$,

$$N_s \ge 2\widetilde{R} \sqrt{\frac{\tau C_\star}{\theta}} \left(\frac{S_{1/2}^{(0)}}{\sqrt{\Delta_s}} + S_{1/2}^{(1)}\right). \tag{102}$$

Squaring equation 102 gives

$$N_s^2 \ge \frac{4\tau C_\star \widetilde{R}^2}{\theta \Delta_s} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{\Delta_s}\right)^2, \tag{103}$$

where we used $(S_{1/2}^{(0)}/\sqrt{\Delta_s} + S_{1/2}^{(1)})^2 = (S_{1/2}^{(0)} + S_{1/2}^{(1)}\sqrt{\Delta_s})^2/\Delta_s$. Substituting the reciprocal of equation 103 into equation 101 gives

$$\begin{aligned} \mathbb{E}F(z_{s+1}) &\le \frac{2\tau C_\star \widetilde{R}^2}{\theta} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{\Delta_s}\right)^2 \\ &\qquad \times \frac{\theta \Delta_s}{4\tau C_\star \widetilde{R}^2 \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{\Delta_s}\right)^2} \\ &= \frac{\Delta_s}{2} = \Delta_{s+1}. \end{aligned} \tag{104}$$

The induction starts from the deterministic identity $F(z_0) = F_0 = \Delta_0$. Equation equation 104 then shows that $\mathbb{E}F(z_s) \le \Delta_s$ implies $\mathbb{E}F(z_{s+1}) \le \Delta_{s+1}$, so the invariant holds for every phase.

Total complexity. Since $S = \lceil \log_2(F_0/\varepsilon) \rceil$, we have $\Delta_S = F_0/2^S \le \varepsilon$. The restart invariant therefore gives $\mathbb{E}[f(z_S) - f(x^*)] = \mathbb{E}F(z_S) \le \varepsilon$.

It remains to account for the ceiling in the phase-length rule. If $S = 0$, the complexity statement is immediate, so suppose $S \ge 1$. We first show that the quantity inside the ceiling is at least one. Applying the coordinate descent argument leading to equation 80 at x^0, separately for every coordinate i, and using f^* as a lower bound on the resulting objective value gives

$$(\nabla_i f(x^0))^2 \le 2\tau C_\star \big(H_{0,i} + H_{1,i} F_0\big) F_0. \tag{105}$$

Indeed, the trial point $x^0 - [\tau C_\star (H_{0,i} + H_{1,i}F_0)]^{-1} \nabla_i f(x^0) e_i$ satisfies the corresponding radius condition by Lemma 10. The coordinate local model and $\tau C_\star (H_{0,i} + H_{1,i}F_0) \ge 2(H_{0,i} + H_{1,i}F_0)$ therefore give

$$\begin{aligned} &f\left(x^0 - \frac{\nabla_i f(x^0)}{\tau C_\star (H_{0,i} + H_{1,i}F_0)} e_i\right) \\ &\qquad \le f(x^0) - \frac{(\nabla_i f(x^0))^2}{2\tau C_\star (H_{0,i} + H_{1,i}F_0)}. \end{aligned}$$

The left-hand side is at least f^*. Subtracting f^*, multiplying by the positive denominator, and rearranging gives equation 105. Summing equation 105 over the coordinates and taking square roots gives

$$\begin{aligned} \big\|\nabla f(x^0)\big\| &\le \sqrt{2\tau C_\star F_0} \sqrt{\sum_{i=1}^d H_{0,i} + F_0 \sum_{i=1}^d H_{1,i}} \\ &\le \sqrt{2\tau C_\star F_0} \left(S_{1/2}^{(0)} + S_{1/2}^{(1)} \sqrt{F_0}\right). \end{aligned} \tag{106}$$

The second inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\sqrt{\sum_i H_{j,i}} \le \sum_i \sqrt{H_{j,i}} = S_{1/2}^{(j)}$. Convexity at x^0, followed by Cauchy–Schwarz and $\big\|x^0 - x^*\big\| \le \widetilde{R}$, gives

$$F_0 \le \big\langle \nabla f(x^0), x^0 - x^* \big\rangle \le \widetilde{R} \big\|\nabla f(x^0)\big\|.$$

Combining this inequality with equation 106 and dividing by $F_0 > 0$ yields

$$\widetilde{R} \left(\frac{S_{1/2}^{(0)}}{\sqrt{F_0}} + S_{1/2}^{(1)}\right) \ge \frac{1}{\sqrt{2\tau C_\star}}. \tag{107}$$

For brevity in the final summation only, let

$$T_s := 2\widetilde{R} \sqrt{\frac{\tau C_\star}{\theta}} \left(\frac{S_{1/2}^{(0)}}{\sqrt{\Delta_s}} + S_{1/2}^{(1)}\right).$$

Since $\Delta_s \le F_0$, equations equation 107 and $\theta \le 1$ imply $T_s \ge T_0 \ge \sqrt{2/\theta} \ge 1$. Hence $N_s = \lceil T_s \rceil \le T_s + 1 \le 2T_s$, and

$$\begin{aligned} N_{\mathrm{tot}}^{\mathrm{nonunif}} &= \sum_{s=0}^{S-1} N_s \\ &\le 4\widetilde{R} \sqrt{\frac{\tau C_\star}{\theta}} \left(S_{1/2}^{(0)} \sum_{s=0}^{S-1} \frac{1}{\sqrt{\Delta_s}} + S S_{1/2}^{(1)}\right). \end{aligned} \tag{108}$$

Since $\Delta_s = F_0/2^s$, the first sum in equation 108 is geometric:

$$\begin{aligned} \sum_{s=0}^{S-1} \frac{1}{\sqrt{\Delta_s}} &= \frac{1}{\sqrt{F_0}} \sum_{s=0}^{S-1} 2^{s/2} \\ &= \frac{2^{S/2} - 1}{(\sqrt{2} - 1)\sqrt{F_0}} \\ &\le \frac{\sqrt{2}}{\sqrt{2} - 1} \frac{1}{\sqrt{\varepsilon}}. \end{aligned} \tag{109}$$

The last inequality in equation 109 uses $2^S \le 2F_0/\varepsilon$ when $S > 0$; when $S = 0$, the sum is empty. Moreover, $S \le \log_2(2F_0/\varepsilon)$, so, with the usual convention that a logarithmic complexity factor is truncated below by one, $S = O(\log(F_0/\varepsilon))$. Substituting these bounds into equation 108 gives

$$N_{\mathrm{tot}}^{\mathrm{nonunif}} = O\left(\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\left(\frac{S_{1/2}^{(0)}}{\sqrt{\varepsilon}} + S_{1/2}^{(1)}\log\frac{F_0}{\varepsilon}\right)\right).$$

□

F Additional Experiments

This section complements the controlled experiment in Section 5. We first compare the full-gradient methods on a large-scale finite-sum classification problem and then isolate the benefit of non-uniform coordinate sampling on an anisotropic objective.

F.1 Full-gradient finite-sum classification

We evaluate whether the accelerated behavior observed on the synthetic chain-cosh objective persists on a large-scale learning problem, while also accounting for the extra oracle work incurred by segment relaxation.

Problem setup. We use all $n = 581{,}012$ examples and 54 features of the scaled `covtype.binary` data from LIBSVM (Chang & Lin, 2011). We append an intercept and rescale every row by the same factor so that $\max_j \|a_j\| = 1$. With signed margin m, we consider the convex C^2 loss

$$\ell(m) := \begin{cases} e^{-m}, & m \ge 0, \\ 1 - m + m^2/2, & m < 0, \end{cases}$$

and solve the regularized finite-sum problem

$$\min_{x\in\mathbb{R}^{55}} f(x) := \frac{1}{n}\sum_{j=1}^{n} \ell(y_j a_j^\top x) + \frac{\lambda}{2}\|x\|^2, \qquad \lambda := 10^{-2}. \tag{110}$$

Since $\ell''(m) \le \ell(m)$ and $\|a_j\| \le 1$, $\left\|\nabla^2 f(x)\right\| \le \lambda + n^{-1}\sum_j \ell(y_j a_j^\top x) \le \lambda + f(x)$. Hence equation 110 is (H_0, H_1)-smooth with $H_0 = \lambda + f^*$ and $H_1 = 1$. We compute the reference value f^* using high-accuracy L-BFGS (Liu & Nocedal, 1989); the final gradient norm is $7.1 \cdot 10^{-8}$.

Methods and metrics. We compare four full-gradient methods. *Ours* is the restarted method in Algorithms 1–2 and adapts its curvature to the current gap while enforcing the local-model radius. *GD-Warmup* is the non-accelerated method of Lobanov & Koloskova (2026, Algorithm 2, Appendix C.4). *Frozen-FGM* is the classical two-sequence accelerated method of Nesterov (1983) with the fixed curvature $M_{\mathrm{frozen}} := H_0 + H_1 F_0$; unlike Ours, it has neither gap adaptation nor restarts. *AGMsDR* is Algorithm 1 of Vankov et al. (2025). It shares segment relaxation and accelerated coupling with Ours, but uses the (L_0, L_1)-smooth step and determines its effective curvature from the achieved decrease. Since the continuation in equation 110 is globally smooth, we run AGMsDR with $(L_0, L_1) = (L_{\mathrm{global}}, 0)$, where $L_{\mathrm{global}} := \sup_x \left\|\nabla^2 f(x)\right\|$. The construction satisfies $L_{\mathrm{global}} \le M_{\mathrm{frozen}}$, so the two classical-smoothness baselines are run under their respective assumptions. All methods start from the same point and run until $F(x)/F_0 \le 10^{-4}$. Figure 2 reports outer iterations, total first-order oracle calls, and wall-clock time. For Ours and AGMsDR, the oracle count includes update and segment-relaxation gradients. Runtime is the median and interquartile range over ten runs after one CUDA warm-up.

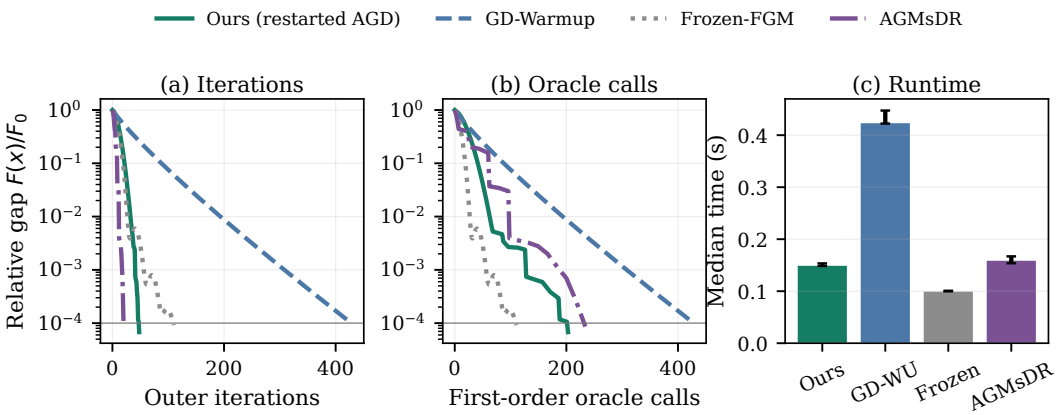

Figure 2: Full-gradient methods on the capped-exponential `covtype.binary` objective: relative gap versus (a) outer iterations and (b) total first-order oracle calls, and (c) runtime. The horizontal line marks the desired relative gap 10^{-4}; runtime bars show medians and interquartile ranges over ten runs on an NVIDIA A100 80GB GPU.

Discussion. The initial portions of the convergence curves in Figures 2(a)–(b) are approximately linear on the semilogarithmic scale. This behavior is consistent with the gap-dominated regime $H_1 F(x) \gg H_0$: both the non-accelerated analysis of GD-Warmup and our accelerated analysis contain a logarithmic dependence on the desired accuracy, corresponding to geometric gap reduction across phases. The distinction is in the dependence on the curvature contribution: our accelerated bound replaces the linear $H_1 \widetilde{R}^2$ factor by its square root. As the iterates approach the optimum and the baseline term H_0 becomes dominant, a globally linear trajectory is no longer predicted.

Our method reaches the desired gap in 48 outer iterations, compared with 110 for Frozen-FGM and 429 for GD-Warmup. AGMsDR requires only 20 outer iterations. Because both accelerated relaxation methods perform additional one-dimensional searches, outer iterations do not tell the whole computational story: Ours uses 48 update and 155 relaxation gradients, for 203 first-order calls in total, whereas AGMsDR uses 20 update and 214 relaxation gradients, for a total of 234. Frozen-FGM uses only 110 first-order calls and has the lowest median runtime, while Ours remains slightly faster than AGMsDR in this implementation. Importantly, AGMsDR is itself an accelerated method for convex (L_0, L_1)-smooth optimization (Vankov et al., 2025). As emphasized in Section 5, we do not claim that Ours uniformly outperforms AGMsDR: the methods address different smoothness models, and this instance indeed favors AGMsDR in outer iterations. Rather, the experiment supports the predicted acceleration under the broader (H_0, H_1)-smoothness model and shows that it can be achieved with competitive total oracle cost.

F.2 Uniform versus non-uniform coordinate sampling

We next isolate the effect of the sampling distribution by changing only the coordinate probabilities while keeping the restart wrapper and accelerated updates fixed. A controlled heterogeneous instance makes the improvement predicted by the coordinate-wise complexity directly measurable.

Problem setup. We consider the separable objective $f(x) := \sum_{i=1}^{24} \phi_i(x_i)$, where, inside a prescribed cutoff, $\phi_i(t) := (H_{0,i}/H_{1,i})(\cosh(\sqrt{H_{1,i}}t) - 1)$, followed by a convex C^2 quadratic continuation. The first four coordinates have $(H_{0,i}, H_{1,i}) = (1, 2)$, while the remaining twenty have $(H_{0,i}, H_{1,i}) = (10^{-2}, 2 \cdot 10^{-2})$. Consequently, $S^{(0)}_{1/2} = 6$ and $S^{(1)}_{1/2} = 6\sqrt{2}$, whereas the corresponding uniform quantities are $d\sqrt{H_0} = 24$ and $d\sqrt{H_1} = 24\sqrt{2}$. This controlled anisotropy isolates the gain predicted by Theorem 8.

Methods and metrics. We compare our uniform and non-uniform coordinate methods from Algorithms 3 and 4. They use the same restart wrapper and accelerated updates; the uniform method samples i_k with probability $1/d$, whereas the non-uniform method uses $p_{k,i} \propto \sqrt{H_{0,i} + H_{1,i} F(y_k)}$

and adjusts its curvature and accelerated weights accordingly. The latter rule extends the classical square-root importance sampling of Nesterov & Stich (2017) by adapting the probabilities to the current gap. Both methods use the same initial points and desired relative gap 10^{-5}. Figure 3 shows pointwise medians and interquartile ranges over 20 random seeds. The left panel counts the coordinate derivatives used in the updates. We implement segment relaxation by golden-section search, which compares only function values and never evaluates a full gradient, using tolerance 10^{-10} and at most 24 search iterations; the right panel adds these function-value queries. Thus the experiment isolates the effect of the sampling rule rather than benchmarking methods based on global coordinate smoothness.

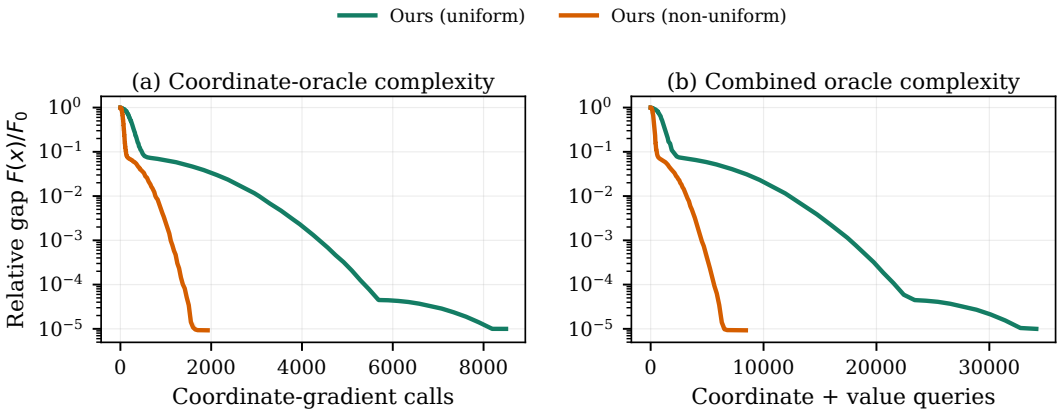

Figure 3: Our uniform and non-uniform coordinate methods on the anisotropic capped-cosh objective: relative gap versus (a) coordinate-gradient calls and (b) coordinate-gradient plus function-value calls. Segment relaxation uses only function comparisons. Curves and bands show medians and interquartile ranges over 20 seeds.

Discussion. Both curves exhibit an approximately linear initial segment on the semilogarithmic scale, again matching the logarithmic, gap-dominated part of the theory. Non-uniform sampling reaches the same decrease using substantially fewer oracle calls. At $F(x)/F_0 \le 10^{-5}$, it reduces the median number of coordinate-gradient calls from 8170 to 1598, a factor of 5.1, and the median combined number of coordinate-gradient and function-value queries from 32,914 to 6453, also a factor of 5.1. No full gradients are used. These gains are consistent with the fourfold reduction of both $S_{1/2}^{(0)}$ and $S_{1/2}^{(1)}$ relative to their uniform upper bounds. The narrow interquartile bands further indicate that the gain is stable across the sampled coordinate sequences. These results support the coordinate-wise complexity in Theorem 8 and show that its dependence on coordinate heterogeneity is visible in practice.

F.3 Inexact Segment-Relaxation Sweep

As noted in Section 6, exact segment minimization in line 3 of Algorithm 2 is a convenient theoretical primitive, but a practical implementation must stop its one-dimensional solver after finitely many oracle queries. This experiment quantifies the resulting tradeoff. We replace the exact minimizer by the verifiable residual from Definition 12, vary its tolerance over five orders of magnitude, and record both the outer convergence and the full cost of the segment solver. This directly tests the inexact-relaxation theory of Appendix G.1: we evaluate the potential inequality in Lemma 13 at every iteration.

Problem setup. We use the chain-cosh objective from Section 5. In dimension $d = 32$, it is

$$f(x) = \frac{\mu_q}{8} x^\top Q x + \psi(u^\top x), \qquad Q = \mathrm{tridiag}(-1, 2, -1), \qquad \|u\| = 1, \tag{111}$$

with $\mu_q = \mu_\psi = 1/2$. The tridiagonal quadratic gives a coupled, ill-conditioned geometry, while the nonlinear term has gap-dependent curvature. For $H_1 = 4$ and $|t| \le c$, we set

$$\psi(t) = \frac{\mu_\psi}{H_1}\big(\cosh(\sqrt{H_1}t) - 1\big) = \frac{1}{8}\big(\cosh(2t) - 1\big). \tag{112}$$

For each initial point, with $F_0 = f(x^0) - f^*$, the cutoff is

$$c = \frac{1}{2}\operatorname{arcosh}(1 + 8F_0)\,. \tag{113}$$

Outside $[-c, c]$, the nonlinear term is continued by

$$\psi_c(t) = \psi(c) + \psi'(c)(|t| - c) + \frac{1}{2}\psi''(c)(|t| - c)^2. \tag{114}$$

The continuation matches the value, first derivative, and second derivative at the cutoff. It is convex and globally smooth, and it preserves $\left\|\nabla^2 f(x)\right\| \le 1 + 4F(x)$. Hence the constants supplied to the proposed method are $H_0 = 1$ and $H_1 = 4$. The unique minimizer is $x^* = 0$, with $f^* = 0$.

Inexact segment routine. At iteration k, define the segment direction and its one-dimensional directional derivative by

$$d_k := x_k - v_k, \qquad y_k(\beta) := v_k + \beta d_k, \qquad g_k(\beta) := \langle \nabla f(y_k(\beta)), d_k \rangle\,, \quad \beta \in [0, 1]. \tag{115}$$

Convexity makes g_k nondecreasing. If $g_k(0) \ge 0$, then v_k minimizes the segment; if $g_k(1) \le 0$, then x_k does. Otherwise an interior zero is bracketed by points $\beta^- < \beta^+$ satisfying $g_k(\beta^-) < 0 \le g_k(\beta^+)$. Derivative bisection repeatedly halves this bracket and returns its upper endpoint as soon as

$$0 \le g_k(\beta^+) \le \delta \qquad \text{and} \qquad f(y_k(\beta^+)) \le f(x_k). \tag{116}$$

The second condition is checked explicitly by function-value queries. To see why the first condition gives the required residual on the entire segment, write $w = v_k + \alpha d_k$, where $\alpha \in [0, 1]$. Then

$$\begin{aligned} \left\langle \nabla f(y_k(\beta^+)), w - y_k(\beta^+) \right\rangle &= (\alpha - \beta^+) g_k(\beta^+) \\ &\ge -\beta^+ g_k(\beta^+) \\ &\ge -\delta. \end{aligned} \tag{117}$$

The first inequality uses $\alpha \ge 0$ and $g_k(\beta^+) \ge 0$; the second uses $\beta^+ \le 1$ and $g_k(\beta^+) \le \delta$. Thus the returned point satisfies both conditions in Definition 12. Each evaluation of g_k uses one full gradient. The implementation uses interval tolerance 10^{-14} as a numerical safeguard and at most 64 bisection iterations; a call that reaches the limit without satisfying equation 116 is recorded as failed.

Method and metrics. We run the restarted full-gradient method from Algorithms 1–2. All trials use $\theta = \tau = 1$ and the same local curvature rule

$$M_k = \max\left\{ 2\big(H_0 + H_1 F(y_k)\big), \frac{\|\nabla f(y_k)\|}{r_1} \right\}. \tag{118}$$

Here

$$r_1 = \frac{1}{(2\sqrt{3} + \sqrt{6})\sqrt{H_1}}$$

is the local-model radius from Algorithm 2. The chain quadratic is strongly convex with certified modulus

$$\mu_{\text{chain}} = \frac{\mu_q}{4}\left(2 - 2\cos\frac{\pi}{d+1}\right).$$

Since the nonlinear term is convex, the initial sublevel set therefore satisfies

$$\left\|x^0 - x^*\right\| \le \widetilde{R} := \sqrt{\frac{2F_0}{\mu_{\text{chain}}}}.$$

This certified value is supplied to the restart wrapper when it computes the phase lengths; the adaptive alternative in Section G.2 is not used in this sweep.

The initial points are sampled as $x^0 \sim \mathcal{N}(0, 0.7^2 I)$, and every run stops when $F(x)/F_0 \le 10^{-6}$ or after 5000 outer iterations. The only quantity varied in the sweep is the fixed absolute segment tolerance

$$\delta \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}\}. \tag{119}$$

Using a fixed δ deliberately isolates its computational effect. A theorem-preserving implementation may instead use the phase-wise schedule in Corollary 14.

For every outer iteration, we compute the accelerated potential

$$\Phi_k := A_k F(x_k) + \frac{1}{2} \|v_k - x^*\|^2 . \tag{120}$$

Lemma 13 predicts the one-step bound

$$\Phi_{k+1} \le \Phi_k + A_{k+1}\delta. \tag{121}$$

We report three diagnostics for this bound. The *potential error* is the sum of positive phase-wise drifts,

$$\sum_s \big(\Phi_{s,\mathrm{end}} - \Phi_{s,0}\big)_+.$$

The *allowance* is the accumulated theoretical error $\sum_{s,k} A_{k+1}\delta$. Finally, the *maximum violation* is

$$\max_{s,k} \big(\Phi_{k+1} - \Phi_k - A_{k+1}\delta\big)_+,$$

which would be positive if the observed one-step potential increase exceeded the allowance in equation 121.

The cost columns distinguish the outer algorithm from its one-dimensional solver. A *relaxation query* is a gradient or function-value query made inside derivative bisection. A *total query* is any full-gradient or function-value query made either by the outer update or by the segment solver. Thus relaxation queries form a subset of total queries and are not counted twice. Runtime is measured end to end. All table entries are medians over five seeded initial points.

Table 2: Inexact-relaxation sweep. Entries are medians over five seeds; all runs reach the target relative gap 10^{-6}.

δ	Outer iter.	Relax. queries	Total queries	Runtime (s)	Max. residual	Potential error	Allowance	Max. violation
10^{-2}	942	1878	3762	0.092	9.98×10^{-3}	596.7	3.466×10^{5}	0
10^{-4}	193	462	848	0.023	9.95×10^{-5}	0	27.39	0
10^{-6}	76	548	700	0.011	9.63×10^{-7}	0	1.647×10^{-2}	0
10^{-8}	77	896	1048	0.015	9.83×10^{-9}	0	1.647×10^{-4}	0
10^{-10}	80	1200	1344	0.018	9.82×10^{-11}	0	1.680×10^{-6}	0

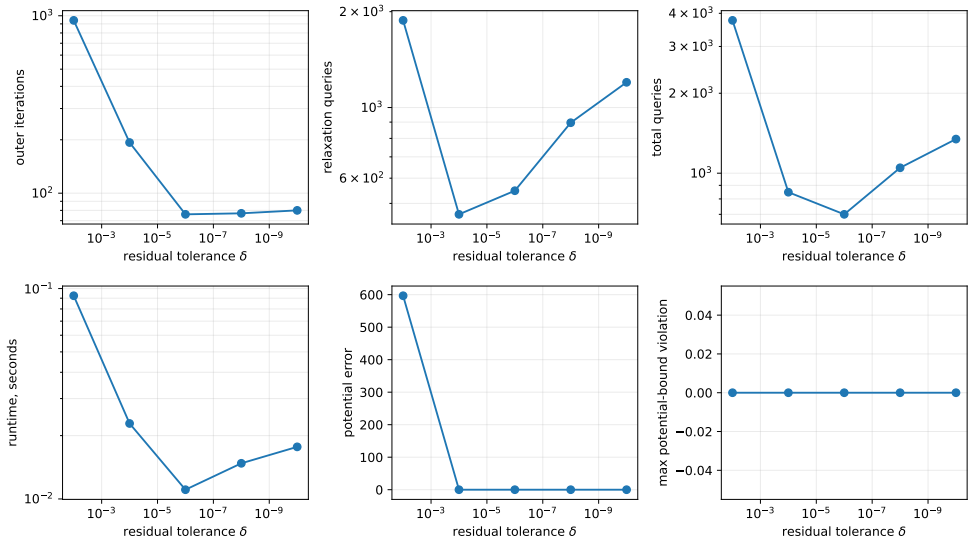

Figure 4: Iteration, primitive-query, runtime, and potential diagnostics as a function of the requested absolute segment residual. The horizontal order is reversed so that increasingly accurate relaxations appear from left to right.

Results. All 25 trials attain the desired relative gap. At the loosest tolerance, $\delta = 10^{-2}$, a segment call needs only about two internal queries on average, but the inaccurate coupling point leads to 942 outer iterations and 3762 total queries. The phase potentials exhibit a positive drift of 596.7. This drift does not contradict equation 121: the corresponding accumulated allowance is 3.466×10^5, so the bound is valid but too loose to be informative about practical progress.

Reducing the tolerance to 10^{-4} lowers the total cost to 848 queries. The smallest observed cost occurs at $\delta = 10^{-6}$: the method needs 76 outer iterations, 548 relaxation queries, and 700 total queries, with median runtime 0.011 seconds. At this tolerance the relaxation is accurate enough that no positive phase-wise potential drift is observed, while the inner search remains inexpensive.

Further tightening does not improve the outer convergence. The counts for $\delta = 10^{-6}, 10^{-8}, 10^{-10}$ are $76, 77, 80$, respectively. In contrast, their relaxation-query counts increase from 548 to 896 and then 1200. Consequently, the nearly exact choice $\delta = 10^{-10}$ uses $1344/700 \approx 1.92$ times as many total queries as $\delta = 10^{-6}$. The experiment therefore exhibits a U-shaped cost–accuracy tradeoff: a loose residual degrades the outer steps, whereas an unnecessarily tight residual spends additional oracle calls without reducing the number of outer iterations.

The diagnostic columns verify that the requested tolerance is actually enforced. For every row, the largest measured residual is below δ; no segment call reaches the 64-iteration limit, and the number of failed relaxations is zero. The maximum one-step potential violation is also zero at machine precision for every tolerance. This is a numerical consistency check of Lemma 13, rather than an independent proof of the lemma.

Discussion. These results show that exact segment minimization is not needed on this problem. A finite residual can preserve the accelerated outer behavior, and the arithmetic spent on the one-dimensional solver should be included when choosing that residual. For the present scaling and target gap, $\delta = 10^{-6}$ gives the best balance between outer and inner work. However, this numerical value is not universal: δ is an absolute directional-derivative tolerance and therefore depends on the scale of the objective, the segment length, and the requested final accuracy. The phase-wise rule in Corollary 14, which controls $\sum_k A_{k+1}\delta_k$, is the appropriate theorem-preserving choice when these scales vary across restart phases. The fixed-tolerance sweep serves a different purpose: it isolates the empirical cost of under-solving and over-solving the same segment subproblem.

Computational environment. All experiments use Python 3.13.11 and NumPy 2.4.1 on an Intel Xeon Gold 6348; the finite-sum runs additionally use PyTorch 2.10.0 and an NVIDIA A100 80GB. The configuration files record all seeds, tolerances, and stopping criteria.

G Practical Extensions

This section records two implementation-oriented extensions of the preceding proofs. Section G.1 replaces exact segment minimization by a verifiable first-order residual and tracks the resulting error through the accelerated potential. Section G.2 then removes prior knowledge of $\widetilde{R}$ from the phase-length rules by restarting unsuccessful trials with doubled budgets. In both cases, we first state the modified guarantee and then derive it using the same notation as in the proofs of Theorems 5– 8. The fixed-residual sweep in Appendix F.3 complements the first extension by measuring its outer convergence, primitive-query cost, and potential error.

G.1 Inexact Segment Relaxation

We begin with the segment step in line 3 of Algorithms 2– 4. Its exact solution enters the proofs through only two consequences: the variational inequality used in the accelerated coupling and the monotonicity relation $f(y_k) \le f(x_k)$ used in the curvature bounds. The following definition relaxes the former while retaining the latter.

Definition 12 (Inexact segment relaxation)**.** Given $v, x \in \mathbb{R}^d$ and $\delta \ge 0$, a point $y \in [v, x]$ is a δ-relaxed segment point if

$$f(y) \le f(x), \qquad \langle \nabla f(y), w - y \rangle \ge -\delta \quad \text{for every } w \in [v, x]. \tag{122}$$

For $\delta = 0$, every exact minimizer over $[v, x]$ satisfies equation 122.

The next lemma isolates the only change in the accelerated potential. It applies to the full-gradient, uniform-coordinate, and non-uniform-coordinate inner methods.

Lemma 13 (Potential stability under inexact relaxation)**.** *At iteration k, replace the exact segment minimizer by a δ_k-relaxed point in the sense of Definition 12. Let*

$$\Phi_k := A_k F(x_k) + \frac{1}{2}\|v_k - x^*\|^2 .$$

For the full-gradient method,

$$\Phi_{k+1} \le \Phi_k + A_{k+1}\delta_k. \tag{123}$$

For either randomized coordinate method, conditionally on the history before the coordinate is sampled,

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k + A_{k+1}\delta_k. \tag{124}$$

Consequently, after N inner iterations,

$$A_N F(x_N) \le \frac{1}{2}\|z - x^*\|^2 + \sum_{k=0}^{N-1} A_{k+1}\delta_k \tag{125}$$

in the full-gradient case. For either coordinate method, the corresponding conditional bound is

$$\mathbb{E}[A_N F(x_N) \mid z] \le \frac{1}{2}\|z - x^*\|^2 + \sum_{k=0}^{N-1} \mathbb{E}[A_{k+1}\delta_k \mid z]. \tag{126}$$

Proof. Fix an iteration k, write $a = a_{k+1}$, and set $A^+ = A_k + a = A_{k+1}$. Define the coupling point

$$w_k := \frac{A_k x_k + a v_k}{A^+} \in [v_k, x_k].$$

The inclusion follows because A_k/A^+ and a/A^+ are nonnegative and sum to one. Convexity of f, applied at y_k with comparison points x_k and x^*, gives

$$\begin{aligned} f(y_k) - f(x_k) &\le \langle \nabla f(y_k), y_k - x_k\rangle , \\ F(y_k) &\le \langle \nabla f(y_k), y_k - x^*\rangle . \end{aligned} \tag{127}$$

The first inequality is the rearrangement of the convexity inequality with comparison point x_k. For the second inequality, convexity gives $f(x^*) \ge f(y_k) + \langle \nabla f(y_k), x^* - y_k\rangle$; rearranging and using $F(y_k) = f(y_k) - f(x^*)$ gives the claimed bound.

Using $A^+ = A_k + a$ and $F(y_k) - F(x_k) = f(y_k) - f(x_k)$, we have

$$\begin{aligned} A^+F(y_k) - A_kF(x_k) &= A_k\big(f(y_k) - f(x_k)\big) + aF(y_k) \\ &\le A_k \langle \nabla f(y_k), y_k - x_k\rangle + a\langle \nabla f(y_k), y_k - x^*\rangle \\ &= \langle \nabla f(y_k), A_k(y_k - x_k) + a(y_k - x^*)\rangle . \end{aligned} \tag{128}$$

The inequality substitutes the two bounds in equation 127; multiplication by the nonnegative weights A_k and a preserves their directions. The final equality uses linearity of the inner product.

The vector in the last line of equation 128 satisfies

$$A_k(y_k - x_k) + a(y_k - x^*) = a(v_k - x^*) + A^+(y_k - w_k). \tag{129}$$

Indeed, substituting $A^+w_k = A_kx_k + av_k$ into the right-hand side gives

$$\begin{aligned} a(v_k - x^*) + A^+(y_k - w_k) &= av_k - ax^* + A^+y_k - A_kx_k - av_k \\ &= A^+y_k - A_kx_k - ax^* \\ &= A_k(y_k - x_k) + a(y_k - x^*), \end{aligned}$$

where the last equality again uses $A^+ = A_k + a$.

Definition 12, applied with $w = w_k$, gives $\langle \nabla f(y_k), w_k - y_k\rangle \ge -\delta_k$. Multiplication by -1 reverses the inequality, and therefore $\langle \nabla f(y_k), y_k - w_k\rangle \le \delta_k$. Substitution of equation 129 into equation 128, followed by this residual bound, yields

$$\begin{aligned} A_{k+1}F(y_k) - A_kF(x_k) &\le a_{k+1}\langle \nabla f(y_k), v_k - x^*\rangle + A_{k+1}\langle \nabla f(y_k), y_k - w_k\rangle \\ &\le a_{k+1}\langle \nabla f(y_k), v_k - x^*\rangle + A_{k+1}\delta_k. \end{aligned} \tag{130}$$

We now insert equation 130 into the potential calculation. For the full-gradient method, the accepted local model gives

$$A_{k+1}F(x_{k+1}) \le A_{k+1}F(y_k) - \frac{A_{k+1}}{2M_k}\left\|\nabla f(y_k)\right\|^2. \tag{131}$$

This is the descent inequality from the exact proof multiplied by the nonnegative weight A_{k+1}. Expanding the squared norm after $v_{k+1} = v_k - a_{k+1}\nabla f(y_k)$ gives

$$\begin{aligned}\frac{1}{2}\left\|v_{k+1} - x^*\right\|^2 = \frac{1}{2}\left\|v_k - x^*\right\|^2 - a_{k+1}\left\langle \nabla f(y_k), v_k - x^*\right\rangle \\ + \frac{a_{k+1}^2}{2}\left\|\nabla f(y_k)\right\|^2.\end{aligned} \tag{132}$$

The middle term in equation 132 is the cross term of the squared norm, and the last term is the squared length of the update. Adding equation 131 and equation 132, then rearranging equation 130 as

$$A_{k+1}F(y_k) - a_{k+1}\left\langle \nabla f(y_k), v_k - x^*\right\rangle \le A_kF(x_k) + A_{k+1}\delta_k,$$

gives

$$\Phi_{k+1} \le \Phi_k + A_{k+1}\delta_k + \left(\frac{a_{k+1}^2}{2} - \frac{A_{k+1}}{2M_k}\right)\left\|\nabla f(y_k)\right\|^2.$$

The weight relation $M_k a_{k+1}^2 = \theta A_{k+1}$ implies

$$\frac{a_{k+1}^2}{2} - \frac{A_{k+1}}{2M_k} = -\frac{(1-\theta)A_{k+1}}{2M_k} \le 0.$$

The equality substitutes $a_{k+1}^2 = \theta A_{k+1}/M_k$; the inequality uses $\theta \in (0,1]$, $A_{k+1} \ge 0$, and $M_k > 0$. Dropping this nonpositive term proves equation 123.

For the uniform-coordinate method, the conditional descent and squared-distance calculations in equation 48 and equation 49 are unaffected by the segment accuracy. Replacing the exact coupling by equation 130 therefore gives

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k + A_{k+1}\delta_k + \left(\frac{da_{k+1}^2}{2} - \frac{A_{k+1}}{2dM_k}\right)\left\|\nabla f(y_k)\right\|^2.$$

Since $d^2 M_k a_{k+1}^2 = \theta A_{k+1}$, the coefficient equals

$$-\frac{(1-\theta)A_{k+1}}{2dM_k} \le 0.$$

Thus dropping the nonpositive gradient term proves equation 124 for uniform sampling.

For non-uniform sampling, the exact calculation equation 90 similarly becomes

$$\begin{aligned}\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k + A_{k+1}\delta_k \\ + \frac{1}{2}\sum_{i=1}^{d}\left(\frac{a_{k+1}^2}{p_{k,i}} - \frac{A_{k+1}p_{k,i}}{M_{k,i}}\right)(\nabla_i f(y_k))^2.\end{aligned}$$

By equation 91, every coefficient in this sum equals $-(1-\theta)A_{k+1}/(\mathcal{M}_k\sqrt{M_{k,i}}) \le 0$. Each squared partial derivative is nonnegative, so the entire sum is nonpositive and may be dropped. This proves equation 124 for non-uniform sampling. In both coordinate cases, y_k, A_{k+1}, and δ_k are $\mathcal{F}_k$-measurable because the segment point and its tolerance are chosen before the coordinate is sampled; hence the residual term remains outside the conditional expectation.

It remains to telescope the one-step estimates. For the full-gradient method, repeated application of equation 123 gives

$$\Phi_N \le \Phi_0 + \sum_{k=0}^{N-1} A_{k+1}\delta_k.$$

Because $A_0 = 0$ and $v_0 = z$, we have $\Phi_0 = \frac{1}{2}\|z - x^*\|^2$. Moreover, $A_N F(x_N) \le \Phi_N$ because the remaining term $\frac{1}{2}\|v_N - x^*\|^2$ is nonnegative. These two observations give equation 125. For a coordinate method, the tower property gives at every iteration

$$\begin{aligned}\mathbb{E}[\Phi_{k+1} \mid z] &= \mathbb{E}[\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \mid z] \\ &\le \mathbb{E}[\Phi_k \mid z] + \mathbb{E}[A_{k+1}\delta_k \mid z].\end{aligned}$$

Summing these inequalities, substituting the same value of Φ_0, and using $A_N F(x_N) \le \Phi_N$ proves equation 126.

Finally, the first condition in equation 122, together with the unchanged descent step, gives $f(x_{k+1}) \le f(y_k) \le f(x_k)$. Thus all local-curvature and radius arguments from the exact proofs continue to hold. □

The lemma controls the accumulated error $\sum_k A_{k+1}\delta_k$. We next convert this global condition into a per-iteration tolerance that is known during a restart phase and show that it preserves the phase contraction.

Corollary 14 (A phase-wise accuracy schedule)**.** *Consider a phase with certified level Δ_s and length N_s. Let $\underline{A}_{s,N_s} > 0$ be a lower bound satisfying $A_{N_s} \ge \underline{A}_{s,N_s}$ for every realization, conditionally on the phase start in the randomized methods. Suppose the inexact points satisfy the weighted schedule*

$$A_{k+1}\delta_{s,k} \le \frac{\underline{A}_{s,N_s}\Delta_s}{4N_s}, \qquad k = 0, \dots, N_s - 1. \tag{133}$$

Then

$$\sum_{k=0}^{N_s-1} A_{k+1}\delta_{s,k} \le \frac{1}{4}\underline{A}_{s,N_s}\Delta_s.$$

If the phase lengths in Theorems 5, 6, and 8 are multiplied by $\sqrt{2}$, the exact part of the one-run estimate is at most $\Delta_s/4$, while the inexactness contributes at most $\Delta_s/4$. Hence every phase still contracts the gap by a factor two, pathwise for the full-gradient method and in expectation for the coordinate methods. The full-gradient and coordinate-gradient complexities stated in the three theorems are therefore unchanged in big-O notation; primitive line-search queries are accounted for separately below.

For the deterministic full-gradient method, the simpler sufficient rule $\delta_{s,k} \le \Delta_s/(4N_s)$ may be used, because $A_{k+1} \le A_{N_s}$ and the pathwise potential can be divided directly by A_{N_s}. The weighted schedule above is used for the randomized methods to avoid dividing an expectation by the random weight A_{N_s}.

Proof. There are N_s terms in the sum. Bounding each term by the right-hand side of equation 133 therefore gives

$$\sum_{k=0}^{N_s-1} A_{k+1}\delta_{s,k} \le N_s \frac{\underline{A}_{s,N_s}\Delta_s}{4N_s} = \frac{1}{4}\underline{A}_{s,N_s}\Delta_s.$$

The equality cancels the positive factor N_s.

For the full-gradient method, substituting this accumulated-error bound into equation 125 and dividing by $A_{N_s} > 0$ gives

$$\begin{aligned}F(x_{N_s}) &\le \frac{\|z_s - x^*\|^2}{2A_{N_s}} + \frac{\underline{A}_{s,N_s}}{A_{N_s}}\frac{\Delta_s}{4} \\ &\le \frac{\|z_s - x^*\|^2}{2\underline{A}_{s,N_s}} + \frac{\Delta_s}{4}.\end{aligned}$$

The second inequality uses $A_{N_s} \ge \underline{A}_{s,N_s} > 0$: replacing A_{N_s} by its lower bound enlarges the first fraction, while $\underline{A}_{s,N_s}/A_{N_s} \le 1$ bounds the second.

For a coordinate method, the lower bound on A_{N_s} holds for every realization conditionally on z_s. Since $F(x_{N_s}) \ge 0$, multiplying it by $F(x_{N_s})$ preserves its direction and gives

$$\underline{A}_{s,N_s}F(x_{N_s}) \le A_{N_s}F(x_{N_s}).$$

Taking conditional expectation, applying equation 126, and using the accumulated error bound pathwise yields

$$\begin{aligned}\underline{A}_{s,N_s}\mathbb{E}[F(x_{N_s}) \mid z_s] &\le \mathbb{E}[A_{N_s}F(x_{N_s}) \mid z_s] \\ &\le \frac{1}{2}\left\|z_s - x^*\right\|^2 + \frac{1}{4}\underline{A}_{s,N_s}\Delta_s.\end{aligned}$$

Factoring $\underline{A}_{s,N_s}$ out of the conditional expectation is valid because it is fixed conditionally on z_s. Dividing by this positive quantity gives

$$\mathbb{E}[F(x_{N_s}) \mid z_s] \le \frac{\left\|z_s - x^*\right\|^2}{2\underline{A}_{s,N_s}} + \frac{\Delta_s}{4}.$$

The exact one-run term is proportional to N_s^{-2}. Multiplication of the exact-segment phase length by $\sqrt{2}$ divides that term by $(\sqrt{2})^2 = 2$, reducing it from $\Delta_s/2$ to $\Delta_s/4$. Adding the inexactness contribution $\Delta_s/4$ gives the required phase bound $\Delta_s/2$.

For completeness, consider the simpler deterministic rule $\delta_{s,k} \le \Delta_s/(4N_s)$. Monotonicity of the accelerated weights gives $A_{k+1} \le A_{N_s}$, and hence

$$\sum_{k=0}^{N_s-1} A_{k+1}\delta_{s,k} \le \sum_{k=0}^{N_s-1} A_{N_s}\frac{\Delta_s}{4N_s} = \frac{1}{4}A_{N_s}\Delta_s.$$

After division by A_{N_s}, this contributes exactly $\Delta_s/4$. For randomized methods the analogous division would place a random A_{N_s} in the denominator of an expectation, which is why the deterministic lower bound $\underline{A}_{s,N_s}$ is used instead. □

Explicit lower bounds. We now identify admissible values of $\underline{A}_{s,N}$ for the three inner methods. The weight estimates already proved for the exact segment step remain valid because the inexact definition preserves $f(x_{k+1}) \le f(y_k) \le f(x_k)$. They give

$$\begin{aligned}\underline{A}_{s,N}^{\mathrm{full}} &:= \frac{\theta N^2}{4\tau C_\star(H_0 + H_1\Delta_s)}, \\ \underline{A}_{s,N}^{\mathrm{coord}}(z_s) &:= \frac{\theta N^2}{4d^2\tau C_\star(H_0 + H_1F(z_s))}, \\ \underline{A}_{s,N}^{\mathrm{nonunif}}(z_s) &:= \frac{\theta N^2}{4\tau C_\star(S_{1/2}^{(0)} + S_{1/2}^{(1)}\sqrt{F(z_s)})^2}.\end{aligned}$$

For the full-gradient method, equation 21 lower bounds A_N by a squared sum of $(H_0 + H_1F(y_k))^{-1/2}$. Monotonicity and the certified phase condition give $F(y_k) \le \Delta_s$; since $H_1 \ge 0$, this implies

$$\frac{1}{\sqrt{H_0 + H_1F(y_k)}} \ge \frac{1}{\sqrt{H_0 + H_1\Delta_s}}.$$

Replacing every one of the N summands by this lower bound and then squaring gives $A_N \ge \underline{A}_{s,N}^{\mathrm{full}}$. The uniform and non-uniform coordinate bounds are exactly equation 58 and equation 96, respectively, with $z = z_s$. Their right-hand sides depend only on the observed phase start, so they are fixed before the coordinate randomness of the inner run is sampled.

At iteration k, a candidate segment point determines its residual and the local model used to compute A_{k+1}. The candidate is accepted only if these two quantities satisfy equation 133; otherwise the one-dimensional search is refined. Thus the schedule can be checked without knowing future accelerated weights.

Primitive line-search cost. The preceding argument counts gradient or coordinate-gradient calls exactly as in the main theorems. We now account separately for the primitive queries needed to enforce the segment residual.

Let $\nu_{s,k}(\delta)$ denote the number of primitive function or directional-derivative queries used to produce a point satisfying Definition 12. Corollary 14 requires residual

$$\delta_{s,k} \le \frac{\underline{A}_{s,N_s}\Delta_s}{4N_sA_{k+1}}.$$

There is one segment search at each inner iteration. Summing its query cost over all iterations and phases therefore gives

$$N_{\mathrm{line}} = \sum_{s} \sum_{k=0}^{N_s-1} \nu_{s,k} \left(\frac{\underline{A}_{s,N_s} \Delta_s}{4 N_s A_{k+1}} \right) \tag{134}$$

primitive line-search queries. Without an additional regularity assumption on the one-dimensional restrictions, this term must be reported separately.

For example, fix $\delta > 0$ and write $\varphi_k(\beta) := f(v_k + \beta(x_k - v_k))$. If φ_k' is $L_{\mathrm{seg},k}$-Lipschitz on $[0,1]$, a derivative bisection first tests whether an endpoint is optimal and otherwise brackets an interior zero β_k^* of φ_k'. If the final bracket has width ρ and β lies in that bracket, Lipschitz continuity and $\varphi_k'(\beta_k^*) = 0$ give

$$|\varphi_k'(\beta)| = |\varphi_k'(\beta) - \varphi_k'(\beta_k^*)| \le L_{\mathrm{seg},k} |\beta - \beta_k^*| \le L_{\mathrm{seg},k} \rho.$$

For $y_k = v_k + \beta(x_k - v_k)$ and $w = v_k + t(x_k - v_k)$, where $t \in [0,1]$, the chain rule gives

$$\langle \nabla f(y_k), w - y_k \rangle = (t - \beta) \varphi_k'(\beta) \ge -|\varphi_k'(\beta)|,$$

because $|t - \beta| \le 1$. Hence $\rho \le \delta / L_{\mathrm{seg},k}$ is sufficient for the variational residual. Starting from an interval of length one, q bisections produce width 2^{-q}. Thus choosing $q \ge \max\{0, \lceil \log_2(L_{\mathrm{seg},k}/\delta) \rceil\}$, together with the endpoint tests, requires

$$O\left(1 + \max\left\{0, \log \frac{L_{\mathrm{seg},k}}{\delta}\right\}\right)$$

directional-derivative queries. The condition $f(y_k) \le f(x_k)$ is checked directly before acceptance. A function-value bracketing method may be used as well if its final bracket can be converted to the same residual using a valid $L_{\mathrm{seg},k}$ bound. Consequently, a logarithmic-cost one-dimensional solver adds only logarithmic factors to equation 134; in coordinate applications, its function evaluations are counted separately from partial derivatives.

G.2 Adaptive Phase Lengths without Prior Knowledge of $\widetilde{R}$

We next address the phase-length rules. Their stated values use $\widetilde{R}$, although the desired phase contraction can be verified from the observable gap because f^* is assumed known. The wrapper below starts from any positive trial budget, reruns every unsuccessful trial from the same phase point, and doubles the budget until the gap is halved. Thus $\widetilde{R}$ is absent from the algorithmic inputs and is used only to bound how soon a trial must succeed.

Algorithm 5 Verified Doubling for One Restart Phase

Input: objective f, known f^*, phase point z_s, certified level Δ_s, inner method $\mathcal{A}$, and initial trial budget $N_0 \ge 1$

$N \leftarrow N_0$
repeat
  Run $\widehat{z} \leftarrow \mathcal{A}(f, z_s, N)$ from the same point z_s
  if $f(\widehat{z}) - f^* \le \Delta_s / 2$ **then**
    return $\widehat{z}$
  end if
  $N \leftarrow 2N$
until the phase is accepted

Return: $z_{s+1} = \widehat{z}$ with $F(z_{s+1}) \le \Delta_s / 2$

Because every trial starts from the same z_s, a failed trial changes neither the certified level Δ_s nor the geometric quantities in the one-run bound. For deterministic full-gradient trials with inexact segment relaxation, we use Corollary 14, with $N_s = N$, or its simpler uniform tolerance.

For the randomized argument below, it is convenient that the entire one-run bound, including segment inexactness, scale as N^{-2}. This is obtained by setting the trial-level error target to

$$\eta_s(N) := \frac{\Delta_s}{8} \left(\frac{N_0}{N} \right)^2$$

and replacing equation 133 by

$$A_{k+1}\delta_{s,k} \le \frac{\underline{A}_{s,N}\eta_s(N)}{N}, \qquad k = 0, \dots, N-1. \tag{135}$$

Summing the N inequalities in equation 135 gives total weighted error at most $\underline{A}_{s,N}\eta_s(N)$. The same division argument as in the proof of Corollary 14 therefore bounds the inexactness contribution by $\eta_s(N)$, which is proportional to N^{-2}. Hence Proposition 16 below applies to exact segment relaxation and to this tightened inexact implementation.

Proposition 15 (Deterministic doubling overhead)**.** *Suppose an inner method deterministically contracts phase s whenever $N \ge N_s^*$. Algorithm 5, started from $N_0 \ge 1$, accepts after fewer than $4\max\{N_0, N_s^*\}$ total inner iterations. Consequently, applying it to the full-gradient method removes $\widetilde{R}$ from the inputs and preserves the complexity of Theorem 5 up to an absolute constant.*

Proof. Let j be the smallest nonnegative integer such that $2^j N_0 \ge N_s^*$. By the assumed contraction guarantee, the trial with budget $2^j N_0$ passes the acceptance test. The preceding trials use budgets $N_0, 2N_0, \dots, 2^{j-1}N_0$, so the geometric-series identity gives

$$\sum_{i=0}^{j} 2^i N_0 = (2^{j+1} - 1)N_0 < 2^{j+1}N_0.$$

If $j = 0$, the total work is N_0, which is smaller than $4\max\{N_0, N_s^*\}$. If $j \ge 1$, minimality of j implies $2^{j-1}N_0 < N_s^*$. Multiplying this strict inequality by four gives $2^{j+1}N_0 < 4N_s^*$. Combining it with the preceding display proves the claimed bound in the remaining case as well. □

Proposition 16 (Randomized coordinate variants)**.** *Suppose that, for every fixed phase start z_s, the conditional one-run estimate of a randomized coordinate inner method is bounded by $C_s(z_s)/N^2$. Let $N_s^*(z_s)$ be large enough that the conditional upper bound satisfies*

$$\frac{C_s(z_s)}{(N_s^*(z_s))^2} \le \frac{\Delta_s}{4}.$$

Use independent coordinate samples in successive trials of Algorithm 5. Then the algorithm terminates almost surely, its accepted point satisfies the phase contraction for every accepted realization, and the expected total number of inner iterations conditionally on z_s is $O(\max\{N_0, N_s^*(z_s)\})$. Hence the uniform- and non-uniform-coordinate complexities in Theorems 6 and 8 are preserved in expectation, while $\widetilde{R}$ is removed from the algorithmic inputs.

Proof. Let $\overline{N}_s$ be the first trial budget satisfying $\overline{N}_s \ge N_s^*(z_s)$. If this is the initial budget, then $\overline{N}_s = N_0$. Otherwise, minimality of the corresponding doubling index gives $\overline{N}_s < 2N_s^*(z_s)$. Hence, in both cases,

$$\overline{N}_s < 2\max\{N_0, N_s^*(z_s)\}. \tag{136}$$

Index the first threshold trial by $j = 0$; its budget is $N_j = 2^j\overline{N}_s$. The assumed N^{-2} estimate and the definition of $N_s^*(z_s)$ imply

$$\begin{aligned} \mathbb{E}[F(\mathcal{A}(f, z_s, N_j)) \mid z_s] &\le \frac{C_s(z_s)}{(2^j\overline{N}_s)^2} \\ &\le \frac{1}{4^j}\frac{C_s(z_s)}{(N_s^*(z_s))^2} \\ &\le \frac{\Delta_s}{4\cdot 4^j}. \end{aligned} \tag{137}$$

The first inequality is the one-run assumption. The second uses $\overline{N}_s \ge N_s^*(z_s)$, and the third is the defining guarantee for $N_s^*(z_s)$. Because $F \ge 0$, Markov's inequality applied to equation 137 gives

$$\begin{aligned} \Pr\left(F(\mathcal{A}(f, z_s, N_j)) > \frac{\Delta_s}{2} \;\middle|\; z_s\right) &\le \frac{\mathbb{E}[F(\mathcal{A}(f, z_s, N_j)) \mid z_s]}{\Delta_s/2} \\ &\le \frac{1}{2\cdot 4^j} =: p_j. \end{aligned}$$

Successive trials use independent randomness conditionally on z_s. Therefore, reaching trial $j \geq 1$ requires the failure of trials $0, \ldots, j-1$, and

$$\begin{aligned} \Pr(\text{reach trial } j \mid z_s) &\leq \prod_{r=0}^{j-1} p_r \\ &= \prod_{r=0}^{j-1} \frac{1}{2 \cdot 4^r} \\ &= 2^{-j} 4^{-j(j-1)/2}. \end{aligned} \tag{138}$$

The last equality uses $\sum_{r=0}^{j-1} r = j(j-1)/2$. Since trial j costs $2^j \overline{N}_s$, the expected work from the first threshold trial onward is bounded by

$$\begin{aligned} \sum_{j=0}^{\infty} 2^j \overline{N}_s \Pr(\text{reach trial } j \mid z_s) &\leq \overline{N}_s \sum_{j=0}^{\infty} 4^{-j(j-1)/2} \\ &= O(\overline{N}_s). \end{aligned} \tag{139}$$

The series is finite because its terms decrease super-geometrically for $j \geq 2$. Moreover, equation 138 tends to zero, so the probability of never accepting is zero. The budgets before $\overline{N}_s$ form a geometric series whose sum is smaller than $\overline{N}_s$. Combining this fact, equation 139, and equation 136 proves

$$\mathbb{E}[N_{s,\text{work}} \mid z_s] = O(\max\{N_0, N_s^*(z_s)\}).$$

The acceptance test is $F(\widehat{z}) \leq \Delta_s/2$, so every returned point satisfies $F(z_{s+1}) \leq \Delta_s/2$.

It remains to compare $N_s^*(z_s)$ with the phase lengths in the main theorems. For the uniform-coordinate method, the one-run estimate equation 59, the radius bound $\|z_s - x^*\| \leq \widetilde{R}$, and the tightened inexact budget above show that the valid choice

$$N_{s,\text{coord}}^*(z_s) := \max\left\{ N_0, \left\lceil 4d\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}}\sqrt{\frac{H_0 + H_1 F(z_s)}{\Delta_s}} \right\rceil \right\} \tag{140}$$

makes the exact one-run term at most $\Delta_s/8$. Indeed, squaring the quantity inside the ceiling gives the lower bound

$$N^2 \geq \frac{16 d^2 \tau C_\star \widetilde{R}^2}{\theta \Delta_s}\big(H_0 + H_1 F(z_s)\big).$$

Taking reciprocals reverses this inequality because both sides are positive. Substitution of the resulting upper bound on $1/N^2$ into equation 59 cancels $d^2 \tau C_\star \widetilde{R}^2 (H_0 + H_1 F(z_s))/\theta$ and leaves $2\Delta_s/16 = \Delta_s/8$. Since $\eta_s(N) \leq \Delta_s/8$ for every trial budget $N \geq N_0$, the exact and inexact contributions sum to at most $\Delta_s/4$, as required in the proposition.

Applying the same calculation to equation 97 gives the valid non-uniform threshold

$$N_{s,\text{nonunif}}^*(z_s) := \max\left\{ N_0, \left\lceil 4\widetilde{R}\sqrt{\frac{\tau C_\star}{\theta}} \frac{S_{1/2}^{(0)} + S_{1/2}^{(1)}\sqrt{F(z_s)}}{\sqrt{\Delta_s}} \right\rceil \right\}. \tag{141}$$

Squaring this threshold and substituting the reciprocal into equation 97 again leaves the exact contribution $2\Delta_s/16 = \Delta_s/8$; adding $\eta_s(N) \leq \Delta_s/8$ gives the required $\Delta_s/4$.

Finally, the verified wrapper gives $F(z_s) \leq \Delta_s$ for every accepted phase start. Because $H_1, S_{1/2}^{(1)} \geq 0$, this pathwise inequality implies

$$\sqrt{\frac{H_0 + H_1 F(z_s)}{\Delta_s}} \leq \sqrt{\frac{H_0}{\Delta_s} + H_1}$$

and

$$\frac{S_{1/2}^{(0)} + S_{1/2}^{(1)}\sqrt{F(z_s)}}{\sqrt{\Delta_s}} \leq \frac{S_{1/2}^{(0)}}{\sqrt{\Delta_s}} + S_{1/2}^{(1)}.$$

Thus equation 140 and equation 141 have the same phase-wise orders as the rules in the main theorems, differing only by absolute constants and ceilings. Summing the conditional expected work over phases and applying the tower property preserves both total expected complexities. □

The doubling construction still relies on exact knowledge of f^* to verify contraction, and it does not remove the need for valid smoothness parameters. It separates two distinct roles of $\widetilde{R}$: the radius continues to describe worst-case geometry in the bounds but need not be known to run the methods.

H Comparison with General Gradient-Dependent ℓ-Smoothness

Following Tyurin (2025), a twice differentiable function is ℓ-smooth if there is a positive, nondecreasing, locally Lipschitz function $\ell : [0, \infty) \to (0, \infty)$ such that

$$\big\|\nabla^2 f(x)\big\| \le \ell\big(\|\nabla f(x)\|\big) \qquad \text{for every } x. \tag{142}$$

For comparison, within the family of convex twice differentiable functions with an attained minimum, let

$$\mathcal{H} := \big\{f : \exists\, 0 \le H_0, H_1 < \infty, \quad \big\|\nabla^2 f(x)\big\| \le H_0 + H_1\big(f(x) - f^*\big) \text{ for all } x\big\},$$

$$\mathcal{L}_{\text{gen}} := \{f : \textit{equation } 142 \text{ holds for some finite-valued positive, nondecreasing, locally Lipschitz } \ell\}.$$

The affine choice $\ell(s) = L_0 + L_1 s$ defines a subclass $\mathcal{L}_{\text{aff}}$. As discussed in the main text and witnessed by Examples 3 and 4, $\mathcal{L}_{\text{aff}} \subsetneq \mathcal{H}$. The relation with the full class $\mathcal{L}_{\text{gen}}$ is different: the two general classes are incomparable. We first clarify why the existing examples do not establish this stronger statement and then prove both non-inclusions by explicit convex examples. For twice differentiable convex functions, the Hessian definition of $\mathcal{H}$ corresponds, up to absolute constants, to Assumption 2; see the discussion following that assumption.

H.1 The existing separation examples are ℓ-smooth

Proposition 17 (Examples 3 and 4 under ℓ-smoothness)**.** *For every $p \in (1, 2]$, both ϕ_p and Φ_p are ℓ_p-smooth with*

$$\ell_p(s) := 2 + s^p.$$

Moreover, the auxiliary function of Tyurin (2025), $\psi_p(r) := r^2/[2\ell_p(4r)]$, is strictly increasing for $r > 0$. Consequently, these examples satisfy the general ℓ-smooth assumptions of Tyurin (2025, Theorems 3.2–3.3), even though they do not satisfy any (L_0, L_1)-smoothness condition with finite constants.

Proof. The definition of q_p and the inequality $\exp(-u) \le 1$ for $u \ge 0$ give

$$\phi_p''(x) = q_p(x) \le 1 + (1 + x^2)^{p/2} \le 2 + |x|^p. \tag{143}$$

The last inequality follows from $(1 + x^2)^{p/2} \le 1 + |x|^p$, valid because $p/2 \in (0, 1]$. On the other hand, $q_p(t) \ge 1$, and therefore

$$|\phi_p'(x)| = \left|\int_0^x q_p(t)\, dt\right| \ge |x|. \tag{144}$$

Combining equation 143 and equation 144 yields

$$|\phi_p''(x)| \le 2 + |\phi_p'(x)|^p = \ell_p\big(|\phi_p'(x)|\big),$$

which proves ℓ_p-smoothness of the one-dimensional example.

For the multidimensional example, $\nabla^2 \Phi_p(x) = \operatorname{diag}(q_p(x_1), 1, \ldots, 1)$. Since $q_p \ge 1$, its spectral norm is $q_p(x_1)$. Also $\|\nabla \Phi_p(x)\| \ge |\phi_p'(x_1)|$. Hence

$$\begin{aligned}\big\|\nabla^2 \Phi_p(x)\big\| &= q_p(x_1) \\ &\le 2 + |\phi_p'(x_1)|^p \\ &\le 2 + \|\nabla \Phi_p(x)\|^p = \ell_p\big(\|\nabla \Phi_p(x)\|\big).\end{aligned}$$

This proves ℓ_p-smoothness of Φ_p.

It remains to check the auxiliary condition used by Tyurin (2025). Direct substitution gives

$$\psi_p(r) = \frac{r^2}{2\ell_p(4r)} = \frac{r^2}{4 + 2 \cdot 4^p r^p}.$$

Differentiating this quotient gives

$$\psi_p'(r) = \frac{8r + 2 \cdot 4^p(2-p)r^{p+1}}{(4 + 2 \cdot 4^p r^p)^2} > 0, \qquad r > 0,$$

where positivity uses $p \le 2$. Thus ψ_p is strictly increasing on $(0, \infty)$, completing the proof. □

H.2 Incomparability of the general classes

The preceding proposition shows that failure of (L_0, L_1)-smoothness in Examples 3 and 4 does not imply failure of general ℓ-smoothness. We now give one example for each direction of the non-inclusion.

Proposition 18 (An ℓ-smooth objective outside $\mathcal{H}$)**.** *The function $u : \mathbb{R} \to \mathbb{R}$ defined by*

$$u(x) := e^{x^2} - 1$$

is convex, coercive, infinitely differentiable, and ℓ_u-smooth with $\ell_u(s) := 2 + \frac{3}{2}s^2$, but $u \notin \mathcal{H}$. Hence $\mathcal{L}_{\mathrm{gen}} \not\subseteq \mathcal{H}$.

Proof. The derivatives of u are

$$u'(x) = 2xe^{x^2}, \qquad u''(x) = (2 + 4x^2)e^{x^2} > 0.$$

Thus u is infinitely differentiable and strictly convex. Moreover, $u(0) = u'(0) = 0$ and $u(x) \to \infty$ as $|x| \to \infty$, so it is coercive and has the unique minimizer $x^* = 0$.

Set $t = x^2$. Since $e^t - 1 \le te^t$ for $t \ge 0$, we obtain

$$\begin{aligned} u''(x) &= 2 + 2(e^t - 1) + 4te^t \\ &\le 2 + 6te^t \\ &\le 2 + 6te^{2t} = 2 + \frac{3}{2}|u'(x)|^2 = \ell_u\big(|u'(x)|\big). \end{aligned}$$

The function ℓ_u is finite-valued, positive, nondecreasing, and locally Lipschitz on $[0, \infty)$. Therefore $u \in \mathcal{L}_{\mathrm{gen}}$.

Suppose instead that $u \in \mathcal{H}$. Then some finite $H_0, H_1 \ge 0$ would satisfy

$$(2 + 4t)e^t \le H_0 + H_1(e^t - 1), \qquad t \ge 0.$$

Dividing by e^t gives

$$2 + 4t \le H_0 e^{-t} + H_1(1 - e^{-t}).$$

As $t \to \infty$, the left-hand side diverges while the right-hand side converges to H_1, a contradiction. Thus $u \notin \mathcal{H}$. □

The reverse non-inclusion requires curvature that becomes unbounded while the gradient remains bounded. The following construction provides it.

Proposition 19 (An (H_0, H_1)-smooth objective outside $\mathcal{L}_{\mathrm{gen}}$)**.** *There exists a convex, coercive, infinitely differentiable function $g : \mathbb{R} \to \mathbb{R}$ with the unique minimizer $x^* = 0$ such that*

$$g''(x) \le 3 + g(x) - g^*, \qquad x \in \mathbb{R}, \tag{145}$$

but g is not ℓ-smooth for any finite-valued nondecreasing function $\ell : [0, \infty) \to (0, \infty)$. Consequently, $g \in \mathcal{H} \setminus \mathcal{L}_{\mathrm{gen}}$, and hence $\mathcal{H} \not\subseteq \mathcal{L}_{\mathrm{gen}}$.

Proof. Define

$$a(t) := \sqrt{1 + t^2}, \qquad q(t) := \operatorname{sech}^2(t) + a(t)\exp\big(-a(t)^6 \sin^2(\pi t)\big), \tag{146}$$

and let

$$g(x) := \int_0^x \int_0^s q(t)\, dt\, ds. \tag{147}$$

The function q is infinitely differentiable, even, and strictly positive. The fundamental theorem of calculus therefore gives

$$g'(x) = \int_0^x q(t)\,dt, \qquad g''(x) = q(x) > 0. \tag{148}$$

Thus g is infinitely differentiable and strictly convex. Moreover, $g(0) = g'(0) = 0$, so its unique minimizer is $x^* = 0$ and $g^* = 0$.

We next verify coercivity and the (H_0, H_1) bound. Since $q(t) \geq \operatorname{sech}^2(t)$, while $(\log\cosh t)'' = \operatorname{sech}^2(t)$, and both functions and their first derivatives vanish at zero, convexity of their difference gives

$$g(x) \geq \log\cosh x. \tag{149}$$

Because $\log\cosh x \to \infty$ as $|x| \to \infty$, the function g is coercive.

The inequalities $\exp(-u) \leq 1$, $\operatorname{sech}^2 x \leq 1$, and $\sqrt{1+x^2} \leq 1 + |x|$ imply

$$q(x) \leq 1 + \sqrt{1+x^2} \leq 2 + |x|. \tag{150}$$

Also $\cosh x \geq e^{|x|}/2$, and hence

$$|x| \leq \log\cosh x + \log 2. \tag{151}$$

Combining equation 149– equation 151, and using $\log 2 < 1$, gives

$$g''(x) = q(x) \leq 2 + \log 2 + \log\cosh x \leq 3 + g(x).$$

This proves equation 145. Applying the same Hessian-to-local-model implication used in the proofs of Examples 3 and 4, namely Liu et al. (2025b, Lemma 2), shows that g satisfies Assumption 2 with $H_0 = 3$ and $H_1 = 1$.

It remains to show that the gradient is bounded although the Hessian is not. For $m \geq 1$, set

$$I_m := \left[m - \frac{1}{2}, m + \frac{1}{2}\right], \qquad A_m := \sqrt{1+m^2}.$$

If $t \in I_m$, then elementary comparison of $1 + t^2$ and $1 + m^2$ gives

$$\frac{A_m}{2} \leq a(t) \leq \frac{3A_m}{2}. \tag{152}$$

Writing $t = m + u$, where $|u| \leq 1/2$, concavity of sine on $[0, 1/2]$ and symmetry give

$$|\sin(\pi t)| = |\sin(\pi u)| \geq 2|u|. \tag{153}$$

Equations equation 152 and equation 153 imply $a(t)^6 \sin^2(\pi t) \geq A_m^6 u^2/16$. Therefore

$$\begin{aligned} &\int_{I_m} a(t)\exp\bigl(-a(t)^6\sin^2(\pi t)\bigr)\,dt \\ &\quad\leq \frac{3A_m}{2}\int_{-\infty}^{\infty}\exp\left(-\frac{A_m^6 u^2}{16}\right)du = \frac{6\sqrt{\pi}}{A_m^2}. \end{aligned} \tag{154}$$

The inequality uses the upper bound on $a(t)$, the lower bound on the exponent, and an enlargement of the integration interval; the equality is the Gaussian integral. The spike term is continuous and hence integrable on $[0, 1/2]$. Since $\sum_{m=1}^{\infty} A_m^{-2} = \sum_{m=1}^{\infty}(1+m^2)^{-1} < \infty$, equation equation 154 yields

$$\int_0^{\infty} a(t)\exp\bigl(-a(t)^6\sin^2(\pi t)\bigr)\,dt < \infty.$$

Together with $\int_0^{\infty}\operatorname{sech}^2(t)\,dt = 1$, this proves that there is a finite constant G such that

$$|g'(x)| \leq G, \qquad x \in \mathbb{R}. \tag{155}$$

Here evenness of q extends the estimate from $x \geq 0$ to all real x.

At every integer $n \geq 1$, $\sin(\pi n) = 0$. Hence

$$g''(n) = q(n) = \operatorname{sech}^2(n) + \sqrt{1+n^2} \geq n. \tag{156}$$

Thus $g''(n) \to \infty$, whereas $|g'(n)| \leq G$. If g were ℓ-smooth for a finite-valued nondecreasing ℓ, then

$$g''(n) \leq \ell\bigl(|g'(n)|\bigr) \leq \ell(G) < \infty$$

for every n, contradicting equation 156. Therefore no such ℓ exists. □

Propositions 18 and 19 establish the precise relationship

$$\mathcal{L}_{\mathrm{aff}} \subsetneq \mathcal{H}, \qquad \mathcal{H} \setminus \mathcal{L}_{\mathrm{gen}} \neq \varnothing, \qquad \mathcal{L}_{\mathrm{gen}} \setminus \mathcal{H} \neq \varnothing.$$

Thus (H_0, H_1)-smoothness is strictly weaker than affine (L_0, L_1)-smoothness, whereas the full (H_0, H_1)-smooth and general ℓ-smooth classes are incomparable. General ℓ-smooth acceleration and our native (H_0, H_1) guarantees are therefore complementary rather than mutually subsuming. Our results cover the full gap-dependent class, including objectives that cannot be controlled by any finite-valued function of the gradient norm alone, and provide uniform and non-uniform coordinate extensions in the native (H_0, H_1) parameters.