LOCALIZE, RESTART, ACCELERATE: STOCHASTIC OPTIMIZATION UNDER GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic convex optimization under asymmetric (L_0, L_1)-generalized smoothness, a model motivated by machine-learning objectives whose local curvature may grow with the gradient norm. We assume an unbiased first-order oracle with additive norm-sub-Gaussian noise. Acceleration is difficult in this setting because momentum may enter regions of much larger curvature, while stochastic gradients cannot reliably certify an unrestricted trajectory. We propose ARC-SG, a two-phase accelerated method: Phase I reduces excessively large gradients using a generalized-smoothness-aware stochastic step, then Phase II solves strongly convex proximal subproblems by a restarted accelerated solver confined to certified smoothness balls. Exact proximal points do not increase the gradient norm, allowing these certificates to propagate through the outer loop. The contribution is a query-by-query certified-localization construction with explicit generalized-smoothness factors and a strongly convex restart extension. ARC-SG achieves, with high probability, an accelerated optimization contribution and smooth-subclass-optimal statistical dependence on accuracy, up to logarithmic and generalized-smoothness factors. Its convex accuracy exponents agree with a contemporaneous public stochastic-acceleration result under a broader smoothness and affine-variance model; our distinction is the certified geometry, explicit parameter accounting, and strongly convex guarantee. The results recover classical accelerated stochastic rates when $L_1 = 0$. Experiments on objectives with unbounded gradients illustrate the two-phase mechanism and its finite-budget advantage.

1 INTRODUCTION

We consider the stochastic convex optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\big[f_\xi(x)\big], \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable and is accessed through an unbiased stochastic first-order oracle. Rather than a global Lipschitz constant for ∇f, we study the gradient-dependent curvature observed in modern learning objectives (Zhang et al., 2020b): $\left\|\nabla^2 f(x)\right\| \le L_0 + L_1 \left\|\nabla f(x)\right\|$, for all $x \in \mathbb{R}^d$. This is (L_0, L_1)-smoothness: $L_1 = 0$ recovers classical smoothness, while $L_1 > 0$ permits unbounded curvature and motivates clipped or normalized steps. For twice-differentiable objectives it implies, up to universal rescaling, the asymmetric finite-difference condition used below.

A substantial literature studies generalized smoothness, first for nonconvex optimization (Zhang et al., 2020a; Crawshaw et al., 2022; Chen et al., 2023; Hübler et al., 2024; Wang et al., 2023) and subsequently for convex problems (Koloskova et al., 2023; Li et al., 2023; Gorbunov et al., 2025; Vankov et al., 2025; Lobanov et al., 2024). Stochastic results include adaptive and clipped SGD (Lobanov & Gasnikov, 2025; Gaash et al., 2025), more general ℓ-smooth objectives (Li et al., 2023), and stochastic proximal-point methods (Tovmasyan et al., 2026). Deterministic accelerated methods attain an $O(1/T^2)$ contribution after an L_1-dependent transient (Gorbunov et al., 2025; Vankov et al., 2025); the near-optimal method of Tyurin (2026) removes nonconstant multiplicative L_1R-dependence from the leading high-accuracy term, where R bounds the initial distance to a solution. Under a broader affine-variance model, Yu et al. (2025) give high-probability guarantees with accelerated low-noise dependence; our norm-sub-Gaussian oracle is its additive-only specialization.

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.

Our contribution is not acceleration under generalized smoothness per se, but a parameter-explicit construction that confines every stochastic query to a certified ordinary-smoothness region and extends to strongly convex objectives. The statistical benchmarks are $\Theta(\sigma^2 R^2/\varepsilon^2)$ in the convex case and $\Theta(\sigma^2/(\mu\varepsilon))$ in the μ-strongly convex case (Nemirovsky & Yudin, 1983; Nemirovski et al., 2009; Lan, 2012; Ghadimi & Lan, 2012).

The obstacle is that momentum can enter regions of much larger gradient and curvature. Deterministic acceleration controls this with coupled clipping (Gorbunov et al., 2025) or monotone adaptive steps (Vankov et al., 2025), using exact gradients to certify the trajectory. A stochastic fluctuation can invalidate such a trajectory-wide certificate, whereas repeatedly reconstructing it would require prohibitively accurate gradient estimates.

We address this difficulty with ARC-SG (**A**ccelerated, **R**estarted, **C**ertified-ball method for **S**tochastic optimization under **G**eneralized smoothness). Phase I reduces gradients to the critical scale L_0/L_1; Phase II applies restarted stochastic acceleration to proximal problems inside certified $O(1/L_1)$-balls. A gradient-scale regularization floor contains the exact prox, while proximal monotonicity, transfer bounds, and a travel–halving argument propagate the certificate and ensure termination.

The novelty is this certified-query geometry, explicit (L_0, L_1)-accounting, and the strongly convex extension. The latter achieves the accuracy-dependent term $\widetilde{O}(\sigma^2/(\mu\varepsilon))$, and $L_1 = 0$ recovers classical accelerated stochastic rates. The inner solver uses standard similar-triangles machinery (Nesterov, 1983; 2018; Lan, 2012; Ghadimi & Lan, 2012; 2013; Gorbunov et al., 2020; Sadiev et al., 2023); the outer construction is related to recursive regularization and localized ball methods (Nesterov, 2012; Allen-Zhu, 2018; Foster et al., 2019; Carmon et al., 2020; 2023; Gruntkowska et al., 2025; Li et al., 2026). Our claim concerns the propagated stochastic certificate, not these ingredients in isolation; Appendix Table 3 gives the detailed comparison.

Contributions. The main contributions are the following:

- **Certified two-phase architecture.** We introduce ARC-SG, combining generalized-smoothness-aware gradient stabilization with restarted accelerated solves on certified proximal balls, with every stochastic query confined to a region carrying an explicit ordinary-smoothness certificate.
- **Convex and strongly convex guarantees.** We prove high-probability guarantees for convex and strongly convex objectives that combine accelerated optimization with smooth-subclass-optimal statistical accuracy dependence; when $L_1 = 0$, the bounds recover standard accelerated stochastic rates.
- **Certificate propagation.** We combine gradient–gap localization, monotonicity of the gradient norm at ball-constrained proximal points, certified smoothness on local balls, and a travel–halving argument to control the number of proximal levels. The claim is about this certified stochastic construction, not priority for each ingredient in isolation.
- **Mechanism-level experiments.** Numerical experiments on objectives with unbounded gradients isolate the roles of stabilization, localization, acceleration, and late-stage averaging.

2 Preliminaries

Let $X^\star := \arg\min_{x\in\mathbb{R}^d} f(x) \neq \varnothing$, and $f^\star := \min_{x\in\mathbb{R}^d} f(x)$, and fix $x^\star \in X^\star$. We assume that a distance certificate $R_0 \geq \left\|x^0 - x^\star\right\|$ is available. For every $x \in \mathbb{R}^d$, define

$$\Delta_x := f(x) - f^\star, \qquad G_x := \|\nabla f(x)\|.$$

In particular, $\Delta_{\text{init}} := \Delta_{x^0} = f(x^0) - f^\star$. We use $1/L_1 := +\infty$ when $L_1 = 0$, and for every $a \geq 0$ set $\log_+ a := \log(\max\{e, a\})$. Thus $\log_+ a = \max\{1, \log a\}$ when $a > 0$, while the convention also covers zero-valued certificates. All vector and operator norms and all inner products are Euclidean. For real u, write $\lceil u \rceil_+ := \max\{0, \lceil u \rceil\}$. We assume $R_0 > 0$; if $R_0 = 0$, then $x^0 = x^\star$ and the method returns x^0. Throughout, $\widetilde{O}$ hides only absolute constants and polylogarithmic factors in the displayed dimensionless problem ratios, $1/\alpha$, and a supplied gap certificate; it never hides polynomial dependence.

Assumption 1 (Convexity). *$f(x)$ is convex and differentiable.*

Assumption 2 ((L_0, L_1)-generalized smoothness)**.** *There exist $L_0 > 0$, $L_1 \geq 0$ such that for all $x, y \in \mathbb{R}^d$ satisfying $\|y - x\| \leq 1/L_1$ we have*

$$\|\nabla f(y) - \nabla f(x)\| \leq \left(L_0 + L_1 \|\nabla f(x)\|\right) \|y - x\|. \tag{2}$$

Assumption 2 states that the gradient is locally Lipschitz continuous. The condition is asymmetric because the right-hand side depends on $\|\nabla f(x)\|$, rather than symmetrically on both x and y. This is the local formulation used by Vankov et al. (2025); Gorbunov et al. (2025). When $L_1 = 0$, the restriction on $\|y - x\|$ disappears and Assumption 2 reduces to ordinary global L_0-smoothness. Moreover, if f is twice differentiable and satisfies

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \|\nabla f(x)\| \qquad \text{for all } x \in \mathbb{R}^d,$$

then Assumption 2 follows after a universal rescaling of the constants L_0 and L_1.

Assumption 3 (Sub-Gaussian stochastic oracle)**.** *Let $\mathcal{F}$ contain all randomness observed before an oracle call, and let x be an $\mathcal{F}$-measurable query point. Every fresh sample $g(x, \xi)$ is conditionally unbiased, and its noise $\zeta(x, \xi) := g(x, \xi) - \nabla f(x)$ is conditionally norm-sub-Gaussian with parameter $\sigma \geq 0$: for $\sigma > 0$,*

$$\mathbb{E}[g(x, \xi) \mid \mathcal{F}] = \nabla f(x), \;\; \mathbb{E}\left[\exp\left(\frac{\|\zeta(x, \xi)\|^2}{\sigma^2}\right) \middle| \mathcal{F}\right] \leq e.$$

Samples within each mini-batch are conditionally independent given $\mathcal{F}$. For $\sigma = 0$, we use the convention $\zeta(x, \xi) = 0$ almost surely, so no division by zero is intended.

Assumption 3 covers almost surely bounded noise and implies $\mathbb{E}\|\zeta\|^2 \leq \sigma^2$. Moreover, define the batched estimator used throughout the method by

$$\mathsf{GradEst}(x, B) := \hat{g}(x; B) := \frac{1}{B} \sum_{i=1}^{B} g(x, \xi_i) \tag{3}$$

from B fresh calls at the same $\mathcal{F}$-measurable point x. By vector Hoeffding for norm-sub-Gaussian random vectors (Jin et al., 2019), an absolute constant $c_g > 0$ exists such that, for every $\rho \in (0, 1)$, $\|\hat{g} - \nabla f(x)\| \leq c_g \sigma \sqrt{\log(2/\rho)/B}$ with conditional probability at least $1 - \rho$. For a proximal objective $F(x) = f(x) + \frac{\lambda}{2}\|x - c\|^2$, we use the same batch and write $\mathsf{GradEst}_F(x, B) := \mathsf{GradEst}(x, B) + \lambda(x - c)$.

The following statements are the geometric core of the method: Lemmas 1–2 justify Phase I, whereas Lemmas 3–4 certify the accelerated proximal solves of Phase II. Proofs are in the supplement.

Local descent under generalized smoothness. This analogue of the classical descent lemma drives Phase I and the gradient–gap relation below.

Lemma 1 (Generalized descent)**.** *Under Assumption 2, if $\|y - x\| \leq 1/L_1$, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + (L_0 + L_1 G_x) \|y - x\|^2 / 2.$$

Gradient localization on sublevel sets. The next lemma converts a function-gap bound into bounds on the gradient norm and local curvature $L_0 + L_1 G_x$, identifying when Phase I may stop and certified acceleration may begin.

Lemma 2 (Gradient–gap relation)**.** *Under Assumptions 1–2, for every $x \in \mathbb{R}^d$,*

$$G_x^2 \leq 2(L_0 + L_1 G_x)\Delta_x, \;\; \text{hence } G_x \leq 2L_1\Delta_x + \sqrt{2L_0\Delta_x}.$$

Consequently, on the sublevel set $\{x \in \mathbb{R}^d : \Delta_x \leq \Delta\}$,

$$L_0 + L_1 G_x \leq 2L_0 + 3L_1^2\Delta.$$

When $L_1 > 0$, define the critical gradient and gap scales by $\overline{G} := L_0/L_1$, and $\overline{\Delta} := L_0/L_1^2$. If $\Delta_x \leq \overline{\Delta}$, then

$$G_x \leq (2 + \sqrt{2})\,\overline{G}, \qquad L_0 + L_1 G_x \leq (3 + \sqrt{2})L_0 < 5L_0.$$

Initial-gap certificates. The algorithm does not require the exact value of the actual initial function gap $\Delta_{\text{init}} := \Delta_{x^0} = f(x^0) - f(x^\star)$. It only uses a certified upper bound on it: Δ_{cert} such that $\Delta_{\text{cert}} \geq \Delta_{\text{init}}$. A canonical choice requiring no additional gap information is

$$\Delta_{\text{def}} := \begin{cases} L_0\big(e^{L_1 R_0} - 1 - L_1 R_0\big)/L_1^2, & L_1 > 0, \\ L_0 R_0^2/2, & L_1 = 0, \end{cases} \tag{4}$$

which always satisfies $\Delta_{\text{init}} \leq \Delta_{\text{def}}$. If another certificate Δ_{user} is available, we use $\Delta_{\text{cert}} := \min\{\Delta_{\text{user}}, \Delta_{\text{def}}\}$. Both Δ_{user} and Δ_{def} are upper bounds on Δ_{init}; hence their minimum is also valid: $\Delta_{\text{init}} \leq \Delta_{\text{cert}}$.

Exact proximal points. For the analysis of Phase II, we introduce the exact solution of a ball-constrained proximal subproblem. Given a center $c \in \mathbb{R}^d$, a regularization parameter $\lambda > 0$, and a radius $r \in (0, +\infty]$, define

$$\widehat{x}(c; \lambda, r) := \arg\min_{x \in \mathbb{B}(c,r)} \left\{ f(x) + \lambda \|x - c\|^2 / 2 \right\}, \tag{5}$$

where $\mathbb{B}(c, r) := \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$. The algorithm does not compute this point exactly: R-STM returns an approximate solution of the corresponding proximal subproblem. The exact point $\widehat{x}(c; \lambda, r)$ serves as an analytical reference for establishing containment, gradient monotonicity, and progress across proximal levels. Its key monotonicity properties are summarized in the following lemma.

Lemma 3 (Proximal gradient and distance monotonicity)**.** *Let Assumption 1 hold and let* $\widehat{x} := \widehat{x}(c; \lambda, r)$. *Then* $\|\nabla f(\widehat{x})\| \leq \|\nabla f(c)\|$, *and* $\|\widehat{x} - c\| \leq \|\nabla f(c)\|/\lambda$.

Moreover, for every $u \in X^\star$, $\|\widehat{x} - u\|^2 \leq \|c - u\|^2 - \|\widehat{x} - c\|^2$. *If* $r = +\infty$, *then* $\|\widehat{x} - c\| = \|\nabla f(\widehat{x})\|/\lambda$.

From gradient caps to certified smoothness balls. The generalized-smoothness condition provides a local curvature bound that depends on the gradient norm at the current point. This is not sufficient by itself to apply R-STM, whose analysis requires a single ordinary smoothness constant on the entire region visited by the inner solver. The following lemma converts a certified gradient cap at the center of a sufficiently small ball into a uniform smoothness bound on that ball.

Lemma 4 (Certified smoothness ball)**.** *Assume* $L_1 > 0$, *let* $\widehat{G} \geq \|\nabla f(c)\|$, *and choose* $0 < r \leq 1/(2L_1)$. *For all* $x, y \in \mathbb{B}(c, 2r)$,

$$G_x \leq e\left(\widehat{G} + \frac{L_0}{L_1}\right), \ \|\nabla f(x) - \nabla f(y)\| \leq L_{\text{ball}} \|x - y\|.$$

where $L_{\text{ball}} := 4(L_0 + L_1\widehat{G})$. *Hence* f *is* L_{ball}*-smooth along every segment in* $\mathbb{B}(c, 2r)$. *For* $L_1 = 0$, *the conclusion holds globally with* $L_{\text{ball}} = L_0$.

3 The ARC-SG Method

Effective-parameter convention. Set $\bar{R} := 3R_0$ and $L_{1,\text{eff}} := \max\{L_1, 1/(4\bar{R})\}$. The default certificate in Algorithm 1 is computed with the original L_1; after the displayed assignment, its procedure calls and schedules use $L_1 = L_{1,\text{eff}}$. Assumptions, theorem statements, and the outer strongly convex wrapper retain the original L_1. This substitution is valid because Assumption 2 is monotone in L_1, and $1 + L_1\bar{R}$ changes by at most a factor $5/4$. If the original $L_1 = 0$, the Phase I stopping test fires at $t = 0$ on its good event.

We now describe the method for solving (1). Phase I stabilizes a potentially very large gradient at the critical scale $\overline{G} = L_0/L_{1,\text{eff}}$, and Phase II accelerates localized proximal problems. Algorithm 1 is the sole algorithmic wrapper; its two calls link directly to Procedures 1 and 2 below. Together they form ARC-SG (**A**ccelerated, **R**estarted, **C**ertified-ball method for **S**tochastic optimization under **G**eneralized smoothness). All procedures inherit the global inputs $(R_0, L_0, L_1, \sigma, \varepsilon, \alpha)$ of Algorithm 1.

Algorithm 1: ARC-SG

Input: $x^0, R_0, L_0, L_1, \sigma; \varepsilon > 0$, $\alpha \in (0,1)$; optional Δ_{user}

Set Δ_{cert} by (4);
if Δ_{user} *is supplied* **then**
  $\Delta_{\text{cert}} \leftarrow \min\{\Delta_{\text{user}}, \Delta_{\text{cert}}\}$;
if $\Delta_{\text{cert}} \le \varepsilon$ **then**
  return x^0
$\bar{R} \leftarrow 3R_0$; $L_1 \leftarrow L_{1,\text{eff}} := \max\{L_1, 1/(4\bar{R})\}$;
$x^{\text{I}} \leftarrow$ Phase I$(x^0, \Delta_{\text{cert}}, \alpha/4)$;
return Phase II$(x^{\text{I}}, \varepsilon, \alpha)$;

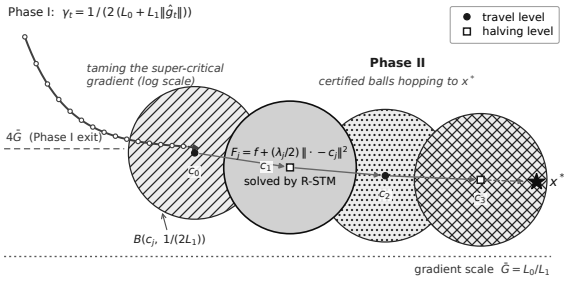


Figure 1: ARC-SG. Phase I stabilizes a super-critical gradient (red); Phase II alternates travel and halving levels inside certified smoothness balls.

Accelerated inner solver. The base STM update is the standard Nesterov estimate-sequence/similar-triangles scheme (Nesterov, 1983; 2018); we do not claim it as a contribution. Procedure S1 states its projected stochastic update. We use it with epoch restarts on shrinking certified balls; the restart schedule and high-probability proof are in Appendix C. Proposition 1 gives $\widetilde{O}(\sqrt{L/\lambda}\log_+(H_0/\delta) + \sigma^2/(\lambda\delta))$ oracle calls, with every query confined to the certified ball.

Procedure S1: STM(F, Q, L, N):

Input: $x_0 = y_0 = z_0 \in Q$, $A_0 = 0$; batch size b
for $t = 0, 1, \ldots, N-1$ **do**
  $a_{t+1} \leftarrow (t+1)/(4L)$; $A_{t+1} \leftarrow A_t + a_{t+1}$;
  $x_{t+1} \leftarrow (A_t y_t + a_{t+1} z_t)/A_{t+1}$;
  $\hat{g}_{t+1} \leftarrow \text{GradEst}_F(x_{t+1}, b)$;
  $z_{t+1} \leftarrow \Pi_Q[z_t - a_{t+1}\hat{g}_{t+1}]$;
  $y_{t+1} \leftarrow (A_t y_t + a_{t+1} z_{t+1})/A_{t+1}$;
return y_N;

Restarted inner solver. Let $c_{\text{stm},1}, c_{\text{stm},2}, c_b > 0$ be absolute concentration constants and set $C_b := 48c_{\text{stm},1}^2 c_b^2 + 70c_b^2 + 1$. The wrapper R-STM$(F, \mathbb{B}(c, r); \lambda, L_F, H_0, \delta, \beta)$ initializes $y^{(0)} = c$ and uses

$$K = \lceil \log_2 \max\{1, H_0/\delta\} \rceil, \quad \bar{\Delta}_k = H_0\, 2^{-k}, \quad N_{\text{ep}} = \lceil 24\sqrt{L_F/\lambda} \rceil,$$
$$Q_k = \mathbb{B}\left(y^{(k)}, 2\sqrt{2\bar{\Delta}_k/\lambda}\right) \cap \mathbb{B}(c, r), \qquad b_k = \max\left\{1, \left\lceil \frac{C_b \sigma^2 \log(8KN_{\text{ep}}/\beta)}{\sqrt{\lambda L_F}\,\bar{\Delta}_k} \right\rceil \right\}.$$

The quantities Q_k, b_k are used only when $K \ge 1$. If $K = 0$, the wrapper returns c; otherwise epoch k runs Procedure S1 from $y^{(k)}$ on Q_k, with batch b_k and failure budget β/K, and returns $y^{(K)}$. Projections are exact and are not counted as stochastic-gradient calls. Appendix C.2 proves the wrapper and gives the constants.

Proposition 1 (Constrained STM and its restarted form)**.** *Let F be convex and L-smooth on a closed convex set Q, let $x_Q \in \arg\min_Q F$, and let $D \ge \max_{x \in Q} \|x - x_Q\|$. With a batch of size b, its noise is ς-norm-sub-Gaussian with $\varsigma \le c_b\sigma/\sqrt{b}$, and Procedure S1 satisfies, with probability at least $1 - \beta$,*

$$F(y_N) - F(x_Q) \le \frac{8LD^2}{N^2} + \frac{c_{\text{stm},1}\varsigma D\sqrt{\log(4/\beta)}}{\sqrt{N}} + \frac{c_{\text{stm},2}\varsigma^2(N+1)\log(4N/\beta)}{L}.$$

If instead F is λ-strongly convex and L_F-smooth on $\mathbb{B}(c, 2r)$, $x_F^\star := \arg\min_{\mathbb{B}(c,r)} F$, and the entry gap satisfies $F(c) - F(x_F^\star) \le H_0$, then R-STM returns y with $F(y) - F(x_F^\star) \le \delta$ and $\|y - x_F^\star\| \le \sqrt{2\delta/\lambda}$ using $\widetilde{O}(\sqrt{L_F/\lambda}\log_+(H_0/\delta) + \sigma^2/(\lambda\delta))$ stochastic gradients; every query lies in $\mathbb{B}(c, 2r)$. The proof and absolute constants are in Appendix C.

Phase-I schedule. With $\bar{G} = L_0/L_1$, define

$$\Lambda_{\text{lg}} := \left\lceil 2 + 2\log_+ \frac{2L_1^2\Delta_{\text{cert}}}{L_0} \right\rceil, \qquad \varepsilon_1 := \frac{\bar{G}}{8\Lambda_{\text{lg}}},$$

$$T_1^{\max} := \max\left\{1, \lceil 6\Lambda_{\lg} L_1 R_0 \rceil,\ 1 + \left\lceil 7 L_1 R_0 \log_+ \frac{2L_1^2 \Delta_{\mathrm{cert}}}{L_0} \right\rceil \right\},$$
$$B_1 := \max\left\{1, \left\lceil \frac{c_g^2 \sigma^2}{\varepsilon_1^2} \log \frac{8(T_1^{\max}+1)}{\alpha} \right\rceil \right\}.$$

What is standard and what is new. Nesterov acceleration, proximal regularization, batching, and restarting are standard ingredients. Our contribution is the certificate system that makes them valid under gradient-dependent curvature: the current gradient cap fixes both the regularization floor and the ball radius; the inner trajectory is confined to that ball; and a transfer estimate plus a fresh measurement propagates the cap to the next level. Thus the smoothness condition is verified for every stochastic query, rather than inferred retrospectively from an unrestricted accelerated path. The same construction supports the outer distance-halving restart used for the strongly convex guarantee.

Procedure 1: Phase I$(x^0, \Delta_{\mathrm{cert}}, \alpha/4)$:

Input: $x_0 = x^0$; $B_1, T_1^{\max}$ as defined above

$\overline{G} \leftarrow L_0/L_1$;
for $t = 0, 1, \ldots, T_1^{\max}$ **do**
  $\widehat{g}_t \leftarrow \mathsf{GradEst}(x_t, B_1)$;
  if $\|\widehat{g}_t\| \le 3\overline{G}$ **then**
    return $x^{\mathrm{I}} \leftarrow x_t$;
  $\gamma_t \leftarrow [2(L_0 + L_1 \|\widehat{g}_t\|)]^{-1}$;
  $x_{t+1} \leftarrow x_t - \gamma_t \widehat{g}_t$;
return $x^{\mathrm{I}} \leftarrow x_{T_1^{\max}+1}$;

Phase I: taming the gradient. Procedure 1 gives the complete stabilization loop. Its step length is at most $1/(2L_1)$, so Lemma 1 always applies; Lemma 2 then controls both descent and the distance to $x^\star$. The certificate Δ_{cert} is used only to set that budget and its confidence allocation; the method never uses the unknown gap Δ_{init}. For an original $L_1 = 0$, the effective-parameter convention makes the initial stopping test fire, so no stabilization update is taken.

Proposition 2 (Phase I guarantee)**.** *Under the effective-parameter convention above, with probability at least $1 - \alpha/4$, Procedure 1 returns x^{I} after at most $1 + 7L_1R_0 \log_+\big(2L_1^2\Delta_{\mathrm{cert}}/L_0\big)$ iterations such that*

$$\big\|\nabla f(x^{\mathrm{I}})\big\| \le 7\overline{G}/2, \qquad f(x^{\mathrm{I}}) \le f(x^0).$$

Every generated iterate x_t, $0 \le t \le T_1^{\max} + 1$, satisfies $\|x_t - x^\star\| \le 3R_0/2$. Its oracle cost is $\widetilde{O}\big(1 + L_1R_0 + (1 + L_1R_0)\sigma^2 L_1^2/L_0^2\big)$. The additive terms are required by the floors $T_1^{\max} \ge 1$ and $B_1 \ge 1$, even when L_1R_0 is small.

Phase-II schedule. Set $\bar{\kappa} := 1 + L_1\bar{R}$, $\log_2^+ a := \max\{0, \log_2 a\}$, and

$$S_G := \max\left\{1, \left\lceil \log_2^+ \frac{10L_0\bar{R}}{L_1\varepsilon} \right\rceil \right\}, \quad S_\Lambda := \max\left\{1, \left\lceil \log_2^+ \frac{128L_0\bar{R}^2}{\varepsilon} \right\rceil \right\},$$
$$J^{\mathrm{sch}} := 60\lceil\bar{\kappa}\rceil S_G(S_\Lambda + 4) + 16, \quad K_{\mathrm{cert}} := \lceil 2\bar{\kappa}(S_G + 4)\rceil + 1, \quad J_{\mathrm{tot}} := J^{\mathrm{sch}} + K_{\mathrm{cert}},$$
$$\iota_{\mathrm{abs}} := \frac{L_0}{8L_1 J_{\mathrm{tot}}}, \qquad \iota_{\mathrm{flr}} := \frac{L_0}{32L_1 S_\Lambda}.$$

For $L_j := 4(L_0 + L_1\hat{G}_j)$, $\lambda_j := \max\{4L_1\hat{G}_j, \Lambda_j\}$, $L_{F,j} := L_j + \lambda_j$, and $H_{0,j} := \hat{G}_j^2/\lambda_j$, use

$$s_j := \begin{cases} \min\{R_0/(4J_{\mathrm{tot}}),\ \iota_{\mathrm{abs}}/L_j\}, & 4L_1\hat{G}_j > \Lambda_j \text{ or finisher}, \\ \min\{R_0/(8S_\Lambda),\ \iota_{\mathrm{flr}}/L_j\}, & 4L_1\hat{G}_j \le \Lambda_j, \end{cases} \qquad \delta_j := \min\{\lambda_j s_j^2/2, \varepsilon/8\}.$$

$$\hat{G}'_{j+1} := \hat{G}_j + L_j s_j, \quad \tau_j := \hat{G}'_{j+1}/16, \quad \beta_j := \frac{\alpha}{4J_{\mathrm{tot}}}, \quad B_j^{\mathrm{est}} := \max\left\{1, \left\lceil \frac{c_g^2\sigma^2}{\tau_j^2} \log \frac{16J_{\mathrm{tot}}}{\alpha} \right\rceil \right\}.$$

After drawing $\hat{g}_j := \mathsf{GradEst}(c_{j+1}, B_j^{\mathrm{est}})$, set $\hat{G}_{j+1} := \min\{\hat{G}'_{j+1}, \|\hat{g}_j\| + \tau_j\}$. For compactness in Procedure 2, write $\Theta_j := (\lambda_j, L_{F,j}, H_{0,j}, \delta_j, \beta_j)$; a level is halving when $\hat{G}_{j+1} \le 3\hat{G}_j/4$. Let $\mathcal{E}$ be the joint event that all scheduled R-STM solves and GradEst bounds hold; then $\mathbb{P}(\mathcal{E}) \ge 1 - \alpha/2$. Appendix E.1 derives the constants; the finisher uses δ_k^{cert} from Procedure 3.

Phase II: accelerated proximal levels. The method maintains a center c_j, a certified gradient cap $\hat{G}_j \geq G_{c_j}$, and a decreasing floor Λ_j. It sets $\lambda_j = \max\{4L_1\hat{G}_j, \Lambda_j\}$, $r_j = 2\hat{G}_j/\lambda_j$, and approximately minimizes $F_j = f + \frac{\lambda_j}{2}\|\cdot - c_j\|^2$ on $\mathbb{B}(c_j, r_j)$. The exact prox lies in the half-radius ball, while every accelerated query remains in the certified smoothness region. A transfer bound propagates the cap, and a fresh GradEst re-anchors it at every level.

For $\hat{x}_j := \widehat{x}(c_j; \lambda_j, r_j)$, Lemmas 3 and 4 give $\|\hat{x}_j - c_j\| \leq r_j/2$ and L_j-smoothness on $\mathbb{B}(c_j, 2r_j)$, so F_j is λ_j-strongly convex and $(L_j+\lambda_j)$-smooth there. A short prox displacement halves the gradient scale; a macroscopic one forces objective decrease, yielding the travel–halving count. The two direct exits certify respectively $f(c_j) - f^\star \leq \varepsilon/2$ and $f(c_{j+1}) - f^\star \leq \Lambda_j\bar{R}^2/2 + \hat{G}_j^2/(2\Lambda_j) + \delta_j \leq 5\varepsilon/8$. The hard schedule cap passes all remaining cases to the certified finisher. Appendix E.2 contains only the detailed inequalities, constants, and proof.

Procedure 2: Phase II$(x^{\mathrm{I}}, \varepsilon, \alpha)$:

Input: $c_1 = x^{\mathrm{I}}$, $\hat{G}_1 = 4L_0/L_1$, $\Lambda_1 = 32L_0$; schedule above

for $j = 1, \ldots, J^{\mathrm{sch}}$ **do**
- **if** $\hat{G}_j \leq \varepsilon/(2\bar{R})$ **then return** c_j;
- **if** $\hat{G}_j < 32\iota_{\mathrm{abs}}$ *or* $j = J^{\mathrm{sch}}$ **then return** Finisher$(c_j, \hat{G}_j)$;
- $L_j \leftarrow 4(L_0 + L_1\hat{G}_j)$;
- $\lambda_j \leftarrow \max\{4L_1\hat{G}_j, \Lambda_j\}$; $r_j \leftarrow 2\hat{G}_j/\lambda_j$;
- $L_{F,j} \leftarrow L_j + \lambda_j$; $H_{0,j} \leftarrow \hat{G}_j^2/\lambda_j$;
- $F_j \leftarrow f + \frac{\lambda_j}{2}\|\cdot - c_j\|^2$;
- $c_{j+1} \leftarrow$ R-STM$(F_j, \mathbb{B}(c_j, r_j); \Theta_j)$;
- $\hat{G}'_{j+1} \leftarrow \hat{G}_j + L_j s_j$;
- $\hat{g}_j \leftarrow$ GradEst$(c_{j+1}, B_j^{\mathrm{est}})$;
- $\hat{G}_{j+1} \leftarrow \min\{\hat{G}'_{j+1}, \|\hat{g}_j\| + \tau_j\}$;
- **if** $\lambda_j = \Lambda_j$ **then**
  - **if** $\Lambda_j \leq \varepsilon/(2\bar{R}^2)$ **then return** c_{j+1};
  - $\Lambda_{j+1} \leftarrow \Lambda_j/2$;
- **else**
  - $\Lambda_{j+1} \leftarrow \Lambda_j$;

Certified finisher. When the gradient cap reaches the drift resolution or the scheduled levels are exhausted, Phase II calls Finisher, stated in Procedure 3. It recomputes the certified accuracy and measurement batch at every regularized ball level, re-anchors the cap with GradEst, and stops on the same gradient certificate. Appendix E.2 proves Lemma 5.

Lemma 5 (Finisher guarantee)**.** *On the joint event $\mathcal{E}$ defined above, from any entry point satisfying $\|\nabla f(c)\| \leq \hat{G} \leq 5\bar{G}$ and $\|c - x^\star\| \leq \bar{R}$, Procedure 3 returns $\hat{x}$ with $f(\hat{x}) - f^\star \leq 7\varepsilon/16 \leq \varepsilon/2$, or exits earlier with gap at most $\varepsilon/2$; it requires no additional random event.*

Proposition 3 (Phase II guarantee)**.** *Suppose* Phase II *is initialized at a point satisfying*

$$\|\nabla f(c_1)\| \leq 4\overline{G}, \qquad \|c_1 - x^\star\| \leq 3R_0/2.$$

With the schedule above, $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha/2$; the finisher adds no new event family. On $\mathcal{E}$, Procedure 2, together with its certified finisher, terminates and returns $\widehat{x}$ with $f(\widehat{x}) - f^\star \leq \varepsilon$. Its oracle cost is

$$\widetilde{O}\Big(L_1R_0 + \bar{\kappa}^{1/2}\sqrt{L_0\bar{R}^2/\varepsilon} + \bar{\kappa}^3\sigma^2\bar{R}^2/\varepsilon^2 + \bar{\kappa}^3\sigma^2L_1^2/L_0^2 \quad\Big),$$

where $\bar{\kappa} = 1 + L_1\bar{R}$ and $\bar{R} = 3R_0$ as defined above.

4 Main Results

Theorem 1 (Convex case)**.** *Let Assumptions 1–3 hold. Fix $x^\star \in X^\star$, a distance certificate $R_0 \geq \|x^0 - x^\star\|$, and a valid initial-gap certificate $\Delta_{\mathrm{cert}} \geq f(x^0) - f^\star$. Let $\varepsilon > 0$, $\alpha \in (0,1)$, and define $\bar{R} := 3R_0$, $\bar{\kappa} := 1 + L_1\bar{R}$. Then for the output $\hat{x}$ of the* ARC-SG*, with probability at least $1 - \alpha$, we have $f(\hat{x}) - f^\star \leq \varepsilon$ after at most*

$$N = \widetilde{O}\Big(\underbrace{1 + L_1R_0}_{\text{Phase I + levels}} + \underbrace{\bar{\kappa}^{1/2}\sqrt{\tfrac{L_0\bar{R}^2}{\varepsilon}}}_{\text{optimization}} + \underbrace{\bar{\kappa}^3\tfrac{\sigma^2\bar{R}^2}{\varepsilon^2}}_{\text{statistical}} + \underbrace{(1 + L_1R_0)\tfrac{\sigma^2L_1^2}{L_0^2} + \bar{\kappa}^3\tfrac{\sigma^2L_1^2}{L_0^2}}_{\text{burn-in}}\Big)$$

stochastic gradient evaluations.

Remark 1 (Default certificate). *The default certificate is always valid: $\Delta_{\mathrm{init}} \le \Delta_{\mathrm{def}}$ (one line, proved in the supplementary material). If ARC-SG runs on $\Delta_{\mathrm{cert}} = \Delta_{\mathrm{def}}$, then with $q := L_1 R_0$ the Phase I entry terms scale as $(1+q)^2$ and the statistical burn-in as $(1+q)^4\sigma^2 L_1^2/L_0^2$: $N = \widetilde{O}\big((1+q)^2 + \bar{\kappa}^{1/2}\sqrt{L_0\bar{R}^2/\varepsilon} + \bar{\kappa}^3\sigma^2\bar{R}^2/\varepsilon^2 + (1+q)^4\sigma^2 L_1^2/L_0^2 + \bar{\kappa}^3\sigma^2 L_1^2/L_0^2\big)$, with $\widetilde{O}$ hiding only absolute constants and polylogarithmic factors in $(q, L_0\bar{R}^2/\varepsilon, 1/\alpha)$. With a user-supplied certificate treated as an independent input, Theorem 1 stands as printed, with its certificate logarithm hidden according to the convention in Section 2.*

Strongly convex restart wrapper. For a known strong-convexity modulus $\mu > 0$, set $x_0 = x^0$ and $K := 1 + \lceil \log_4 \max\{1, \mu R_0^2/(8\varepsilon)\} \rceil$. For $k = 0, \dots, K-1$, let $R_k := R_0 2^{-k}$, $\varepsilon_k := \max\{\mu R_k^2/8, \varepsilon\}$, $\alpha_k := \alpha/K$, and $L_{1,\mathrm{eff},k} := \max\{L_1, 1/(12R_k)\}$. With $\psi(t) := e^t - 1 - t$, set $\Delta_{\mathrm{def},k} := L_0\psi(L_{1,\mathrm{eff},k}R_k)/L_{1,\mathrm{eff},k}^2$. Without a user certificate, use $\Delta_{\mathrm{cert},0} := \Delta_{\mathrm{def},0}$ and, for $k \ge 1$, $\Delta_{\mathrm{cert},k} := \min\{\varepsilon_{k-1}, \Delta_{\mathrm{def},k}\}$. If Δ_{user} is supplied, replace these by $\Delta_{\mathrm{cert},0} := \min\{\Delta_{\mathrm{user}}, \Delta_{\mathrm{def},0}\}$ and, for $k \ge 1$, $\Delta_{\mathrm{cert},k} := \min\{\Delta_{\mathrm{user}}, \varepsilon_{k-1}, \Delta_{\mathrm{def},k}\}$. Set x_{k+1} to the output of ARC-SG run from x_k with $(R_k, \varepsilon_k, \alpha_k, \Delta_{\mathrm{cert},k})$, and return x_K. A supplied Δ_{user} remains valid by induction: a certificate exit leaves the iterate unchanged, whereas an executed round has $\varepsilon_k < \Delta_{\mathrm{user}}$ and returns a point with gap at most ε_k.

Theorem 2 (Strongly convex case). *Under the assumptions and initialization of Theorem 1, suppose additionally that f is μ-strongly convex. Then the restart wrapper above returns $\hat{x}$ such that $f(\hat{x}) - f^\star \le \varepsilon$ with probability at least $1-\alpha$ using*

$$N = \widetilde{O}\Big((1+L_1R_0)\Big(1+\sqrt{\tfrac{L_0}{\mu}}\Big) + \sqrt{\tfrac{L_0}{\mu}}\log_+\!\big(\tfrac{\mu R_0^2}{\varepsilon}\big) + \Big(1+L_1\sqrt{\tfrac{\varepsilon}{\mu}}\Big)^3\tfrac{\sigma^2}{\mu\varepsilon} + \tfrac{\sigma^2 L_1^3 R_0}{\mu^2} + \bar{\kappa}^3\tfrac{\sigma^2 L_1^2}{L_0^2}\Big)$$

oracle calls. In particular, for $\varepsilon \le \mu/(36L_1^2)$ the accuracy-dependent statistical term is $\widetilde{O}(\sigma^2/(\mu\varepsilon))$, whose ε-exponent is optimal on the smooth subclass; the other σ-dependent terms displayed above are ε-independent burn-in.

Remark 2 (Default certificates for the restart). *With a user-supplied certificate, Theorem 2 holds as printed. If every restart uses the default certificate, then, with $q := L_1R_0$, add $\widetilde{O}((1+q)^2 + (1+q)^4\sigma^2L_1^2/L_0^2)$ to its displayed bound. These terms are independent of ε, so the stated $\widetilde{O}(\sigma^2/(\mu\varepsilon))$ accuracy dependence is unchanged.*

For $L_1 = 0$ (via the $L_{1,\mathrm{eff}}$ substitution) both theorems reduce to $\widetilde{O}\big(1 + \sqrt{L_0R_0^2/\varepsilon} + \sigma^2R_0^2/\varepsilon^2\big)$ and $\widetilde{O}\big(\sqrt{L_0/\mu}\,\log_+(\mu R_0^2/\varepsilon) + \sigma^2/(\mu\varepsilon)\big)$—the complexities of accelerated stochastic approximation (Lan, 2012; Ghadimi & Lan, 2013) up to logarithms—and for every $L_1 \ge 0$ the ε-exponents of both the optimization and the statistical terms are unimprovable, since (L_0, L_1)-smooth functions include the L_0-smooth ones and the classical lower bounds (Nemirovsky & Yudin, 1983) apply on that subclass. Whether the multiplicative $\bar{\kappa}$-dependence is necessary remains open.

Proof ideas. The floor $\lambda_j \ge 4L_1\hat{G}_j$ certifies a smoothness ball containing every inner stochastic query, while proximal gradient monotonicity and a transfer bound control the drift of approximate centers. A travel–halving potential bounds the productive levels in each gradient band, and deterministic schedule caps make the confidence union bound finite. Distance-halving restarts then yield the strongly convex result. Complete proofs and constants are given in the supplementary material.

5 Numerical Experiments

We evaluate practical ARC-SG, a computationally streamlined implementation designed to illustrate the two-phase mechanism of the method. The implementation follows the architecture of ARC-SG, but uses less conservative constants and stopping rules; all deviations from the analyzed schedule are listed in the supplementary material.

Experimental setup. We use two certified $d = 40$ families with unbounded gradients and $(L_0, L_1) = (2.1258, 1/\ln 2)$: a convex mixture of cosh-type and ten quartic coordinates, initialized above the critical gradient scale, and a cosh-type objective that is globally strongly convex with $\mu = 0.245$. The oracle adds Gaussian noise satisfying Assumption 3. Curves show the geometric mean over ten independently regenerated instance–noise seeds; bands show one multiplicative standard deviation. The certificate and full protocol are in the supplementary material.

Baselines. We compare with generalized-smoothness SGD (Lobanov & Gasnikov, 2025), clipped SGD (Koloskova et al., 2023; Gaash et al., 2025), fixed-curvature stochastic acceleration, ClipSSTM (Gorbunov et al., 2020), Polyak–Juditsky-averaged SGD, and decreasing-stepsize SGD; all free parameters are selected on separate validation runs.

Procedure 3: Finisher$(c_0, \hat{G}_0)$:

Input: $K_{\mathrm{cert}}, J_{\mathrm{tot}}, \iota_{\mathrm{abs}}$ as defined above

$\lambda_{\mathrm{cert}} \leftarrow L_1\varepsilon/\bar{R}$; $r_{\mathrm{safe}} \leftarrow (2L_1)^{-1}$; $\beta \leftarrow \alpha/(4J_{\mathrm{tot}})$;

for $k = 0, 1, \ldots, K_{\mathrm{cert}} - 1$ **do**

  if $\hat{G}_k \le \varepsilon/(2\bar{R})$ **then return** c_k;

  $L_k \leftarrow 4(L_0 + L_1\hat{G}_k)$;

  $s_k \leftarrow \min\{R_0/(4J_{\mathrm{tot}}), \iota_{\mathrm{abs}}/L_k\}$;

  $\hat{G}'_{k+1} \leftarrow \hat{G}_k + L_k s_k$; $\tau_k \leftarrow \hat{G}'_{k+1}/16$;

  $\delta_k^{\mathrm{cert}} \leftarrow \min\{\varepsilon/(64L_1\bar{R}), \lambda_{\mathrm{cert}} s_k^2/2\}$;

  $H_{0,k}^{\mathrm{cert}} \leftarrow \hat{G}_k/(2L_1)$;

  $B_k^{\mathrm{est}} \leftarrow \max\{1, \lceil c_g^2\sigma^2\tau_k^{-2}\log(16J_{\mathrm{tot}}/\alpha)\rceil\}$;

  $F_k \leftarrow f + \frac{\lambda_{\mathrm{cert}}}{2}\|\cdot - c_k\|^2$;

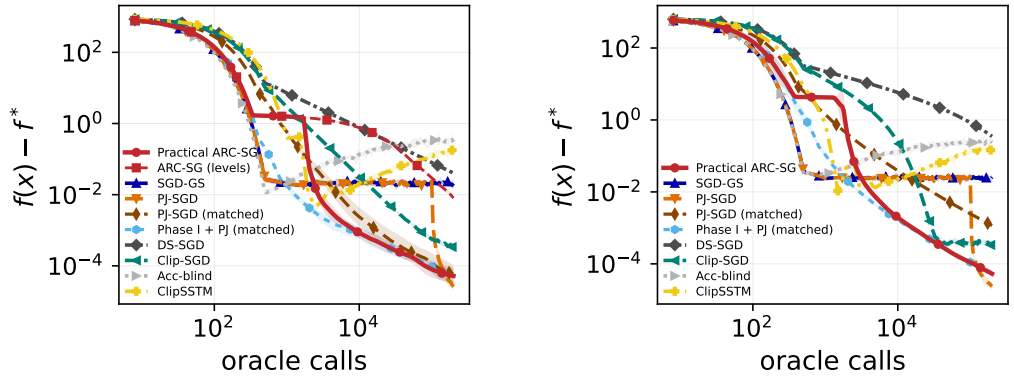
  $c_{k+1} \leftarrow \mathsf{R\text{-}STM}(F_k, \mathbb{B}(c_k, r_{\mathrm{safe}}); \lambda_{\mathrm{cert}}, L_k + \lambda_{\mathrm{cert}}, H_{0,k}^{\mathrm{cert}}, \delta_k^{\mathrm{cert}}, \beta)$;

  $\hat{g}_k \leftarrow \mathsf{GradEst}(c_{k+1}, B_k^{\mathrm{est}})$;

  $\hat{G}_{k+1} \leftarrow \min\{\hat{G}'_{k+1}, \|\hat{g}_k\| + \tau_k\}$;

return $c_{K_{\mathrm{cert}}}$;

Results. Figure 2 illustrates the mechanism rather than the worst-case constants. Practical ARC-SG achieves smaller gaps over most of the budget than methods that are non-accelerated or use a fixed curvature estimate. Ablations show that Phase I is essential on the strongly convex instance, accelerated proximal levels drive the initial drop, and uncovered Polyak–Juditsky averaging supplies much of the late-stage gain. The PJ-SGD baseline reaches a lower gap by the end of the run, whereas practical ARC-SG makes faster progress over most of the finite budget; its observed advantage is therefore speed rather than final accuracy.



(a) Convex, quartic flat directions

(b) μ-strongly convex problem

Figure 2: Objective gap versus the number of stochastic oracle calls. Both axes are logarithmic. Curves show the geometric mean over 10 independently regenerated instance–noise runs; shaded bands indicate one multiplicative standard deviation across runs.

6 Conclusion

We introduced ARC-SG, a high-probability accelerated stochastic method for convex and strongly convex optimization under asymmetric (L_0, L_1)-generalized smoothness. Its central contribution is a parameter-explicit, query-by-query certificate system: gradient-scale regularization confines every accelerated stochastic subproblem to a ball of controlled curvature, while transfer estimates propagate the certificate across proximal levels and restarts. This yields accelerated optimization terms and smooth-subclass-optimal accuracy exponents. The central distinction is the certified geometry and the strongly convex extension. The guarantees require known problem parameters and exact projections, while the experiments use a streamlined variant on synthetic instances; closing this theory–practice gap is the main limitation.

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

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/dba4c1a117472f6aca95211285d0587e-Abstract.html.

Yair Carmon, Arun Jambulapati, Yujia Jin, Yin Tat Lee, Daogao Liu, Aaron Sidford, and Kevin Tian. ReSQueing parallel and private stochastic convex optimization. *arXiv preprint arXiv:2301.00457*, 2023. URL https://arxiv.org/abs/2301.00457.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5396–5427, 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 (NeurIPS)*, volume 35, 2022.

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 32nd Annual Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1319–1345, 2019.

Ofir Gaash, Kfir Yehuda Levy, and Yair Carmon. Convergence of clipped SGD on convex (L_0, L_1)-smooth functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ae6c3f14de6f59ff9a44425a792e5bd1-Abstract-Conference.html.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23 (4):2061–2089, 2013.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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 (ICLR)*, 2025. URL https://arxiv.org/abs/2409.14989.

Kaja Gruntkowska, Hanmin Li, Aadi Rane, and Peter Richtárik. The ball-proximal ("broximal") point method: A new algorithm, convergence theory, and applications. *arXiv preprint arXiv:2502.02002*, 2025. URL https://arxiv.org/abs/2502.02002.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4861–4869, 2024.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363, 2023.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1): 365–397, 2012.

Hanmin Li, Kaja Gruntkowska, and Peter Richtárik. Stabilized proximal point method via trust region control. *arXiv preprint arXiv:2604.02943*, 2026. URL https://arxiv.org/abs/2604.02943.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Aleksandr Lobanov and Anastasia Koloskova. Avoiding bias in clipped SGD for overparameterized models under generalized smoothness. *arXiv preprint arXiv:2605.14800*, 2026. URL https://arxiv.org/abs/2605.14800.

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. URL https://arxiv.org/abs/2412.17050.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3325–3334, 2018.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Yurii 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.

Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, 88: 10–11, 2012.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648, 2023.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Zhirayr Tovmasyan, Grigory Malinovsky, Laurent Condat, and Peter Richtárik. Revisiting stochastic proximal point methods: Generalized smoothness and similarity. *Journal of Nonlinear and Variational Analysis*, 10(3): 471–505, 2026. doi: 10.23952/JNVA.10.2026.3.01. URL https://doi.org/10.23952/JNVA.10.2026.3.01.

Alexander Tyurin. Near-optimal convergence of accelerated gradient methods under generalized and (L_0, L_1)-smoothness. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2508.06884.

Daniil Vankov, Anton Rodomanov, Angelia Nedić, Lalitha Sankar, and Sebastian U. Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/28a9255e2a0e7062e8e0af715f2e5b86-Paper-Conference.pdf.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1195–1204, 2019.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of the 36th Annual Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 161–190, 2023.

Chenhao Yu, Yusu Hong, and Junhong Lin. Convergence analysis of stochastic accelerated gradient methods for generalized smooth optimizations. *arXiv preprint arXiv:2502.11125*, 2025. URL https://arxiv.org/abs/2502.11125.

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 (NeurIPS)*, volume 33, 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 (ICLR)*, 2020b.

APPENDIX ROADMAP

This appendix uses the notation, problem formulation, assumptions, gradient estimator, algorithms, and structural statements of the main paper without restating them. It supplies the omitted proofs, explicit parameter schedules, analyses of the constrained STM routine and the certified finisher stated in the main paper, the strongly convex restart analysis, additional comparisons and extensions, and the full experimental protocol. References to numbered assumptions, lemmas, theorems, procedures, and figures point to their original statements in the main paper.

CONTENTS

A Proofs and auxiliary structural lemmas 13

B Concentration toolbox 16

C Inner stochastic solver: constrained batched STM with epoch restarts 17

D Phase I analysis 19

E Complete ARC-SG schedule and Phase II analysis 21

F Strongly convex restarts 29

G Supplementary remarks and comparisons 31

H Strong-growth and interpolation extensions 33

I Complete experimental details 46

J Additional experiments and limitations 55

A PROOFS AND AUXILIARY STRUCTURAL LEMMAS

Throughout the appendix we use the formulation and notation of Section 2, Assumptions 1–3, and the structural results stated in the main paper. We record below only auxiliary statements needed by the proofs. The first one controls gradient growth along a segment and is used to prove the certified-ball lemma.

Lemma 6 (Gradient growth). *For all x, y with $r = \|y - x\|$, $G_y \le e^{L_1 r} G_x + \frac{L_0}{L_1}(e^{L_1 r} - 1) \le e^{L_1 r}(G_x + L_0 r)$. At $L_1 = 0$ the middle term is read by continuity as $L_0 r$ (since $(e^{L_1 r} - 1)/L_1 \to r$).*

Corollary 1 (Effective smoothness on sublevel sets). *Let $G(\Delta) := 2L_1\Delta + \sqrt{2L_0\Delta}$. On the sublevel set $\{x : \Delta_x \le \Delta\}$ one has $G_x \le G(\Delta)$ and $L_0 + L_1 G_x \le 2L_0 + 3L_1^2\Delta$. In particular, for the* critical gap $\bar{\Delta} := L_0/L_1^2$ *and the* critical gradient $\bar{G} := L_0/L_1$*: on $\{\Delta_x \le \bar{\Delta}\}$ the local smoothness modulus is at most $5L_0$. At $L_1 = 0$ one sets $\bar{\Delta} = \bar{G} = +\infty$; the last claim then asserts modulus at most $5L_0$ on $\{\Delta_x < \infty\}$, where it is L_0.*

Remark 3 ($\mu \le L_0$ in the strongly convex case). *If f is moreover μ-strongly convex, then $\mu \le L_0$: strong convexity with $\nabla f(x^\star) = 0$ gives $f(y) \ge f^\star + \frac{\mu}{2}\|y - x^\star\|^2$ for* all *y, while Lemma 1 at the base $x^\star$ (where $G_{x^\star} = 0$) gives $f(y) \le f^\star + \frac{L_0}{2}\|y - x^\star\|^2$ for $\|y - x^\star\| \le 1/L_1$ (for every y if $L_1 = 0$); applying both at the single point $y = x^\star + tv$ with $\|v\| = 1$ and $t := \min\{1, 1/L_1\} > 0$ yields $\frac{\mu}{2}t^2 \le \frac{L_0}{2}t^2$.*

Lemma 7 (Initial gap). *$\Delta_{\mathrm{init}} \le \Delta_{\mathrm{def}}$, where $\Delta_{\mathrm{def}} := \frac{L_0}{L_1^2}\psi(L_1 R_0)$ with $\psi(t) = e^t - 1 - t \le \frac{t^2}{2}e^t$ for $L_1 > 0$, and $\Delta_{\mathrm{def}} := \frac{L_0 R_0^2}{2}$ at $L_1 = 0$ (the continuous extension, since $\psi(t)/t^2 \to \frac{1}{2}$ as $t \to 0$).*

We use the ball-constrained proximal point $\hat{x}(c; \lambda, r)$ defined in (5) of the main paper. (Throughout, r and its decorated variants denote ball radii; the symbol ρ is reserved for the strong growth constant of Appendix H.)

Lemma 3 follows from monotonicity of ∇f combined with the Karush–Kuhn–Tucker (KKT) stationarity condition $\nabla f(\hat{x}) = -(\lambda + \eta)(\hat{x} - c)$, $\eta \ge 0$: the gradient at the (possibly ball-constrained)

exact prox is *anti-parallel* to the displacement, which forces its norm below $\|\nabla f(c)\|$. Thus the gradient norm at the exact proximal reference point cannot increase. The algorithmic center is only an approximate prox, however, so its certified cap need not be monotone: the transfer estimate gives $\hat{G}^+ \le \hat{G} + L_c s$, and Lemma 13 together with the travel–halving potential controls the accumulated rise.

Finally, we record the smoothness certificate for balls anchored at points with controlled gradients; it follows from Lemma 6 and (2).

A.1 Proofs of the structural lemmas

Throughout this section Assumptions 1 and 2 are in force. Recall $G_x = \|\nabla f(x)\|$, $\Delta_x = f(x) - f^\star$.

A.1.1 Proof of Lemma 6

If $L_1 = 0$ the claim is the definition of L_0-smoothness, so let $L_1 > 0$. Let $r = \|y - x\|$, $m := \max\{1, \lceil L_1 r \rceil\}$ and $u_i := x + \frac{i}{m}(y - x)$, $i = 0, \dots, m$, so that $\|u_{i+1} - u_i\| = r/m \le 1/L_1$. Applying (2) on each segment with base point u_i,

$$G_{u_{i+1}} \le G_{u_i} + \big(L_0 + L_1 G_{u_i}\big)\tfrac{r}{m} = \Big(1 + \tfrac{L_1 r}{m}\Big) G_{u_i} + \tfrac{L_0 r}{m}.$$

Unrolling with $q := 1 + L_1 r/m$,

$$G_y \le q^m G_x + \tfrac{L_0 r}{m} \cdot \tfrac{q^m - 1}{q - 1} = q^m G_x + \tfrac{L_0}{L_1}(q^m - 1) \le e^{L_1 r} G_x + \tfrac{L_0}{L_1}\big(e^{L_1 r} - 1\big),$$

using $q^m = (1 + L_1 r/m)^m \le e^{L_1 r}$. The second form follows from $e^t - 1 \le t e^t$.

□

A.1.2 Proof of Lemma 1

For $t \in [0, 1]$ the point $x_t := x + t(y - x)$ satisfies $\|x_t - x\| \le \|y - x\| \le 1/L_1$, so (2) with base x gives $\|\nabla f(x_t) - \nabla f(x)\| \le (L_0 + L_1 G_x) t \, \|y - x\|$. Hence

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x),\, y - x \rangle &= \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, y - x \rangle \, dt \\ &\le \big(L_0 + L_1 G_x\big) \|y - x\|^2 \int_0^1 t \, dt. \end{aligned}$$

which is the claim.

□

A.1.3 Proof of Lemma 2 and Corollary 1

Let $s := \frac{1}{L_0 + L_1 G_x}$ and $y := x - s\nabla f(x)$; then $\|y - x\| = s G_x = \frac{G_x}{L_0 + L_1 G_x} \le \frac{1}{L_1}$, so Lemma 1 applies:

$$f^\star \le f(y) \le f(x) - s G_x^2 + \tfrac{L_0 + L_1 G_x}{2} s^2 G_x^2 = f(x) - \tfrac{s}{2} G_x^2,$$

i.e., $G_x^2 \le 2(L_0 + L_1 G_x)\Delta_x$, the gradient–gap inequality of Lemma 2. Viewing it as a quadratic inequality in G_x, $G_x \le L_1 \Delta_x + \sqrt{L_1^2 \Delta_x^2 + 2 L_0 \Delta_x} \le 2 L_1 \Delta_x + \sqrt{2 L_0 \Delta_x}$. For the corollary, monotonicity of $G(\cdot)$ gives $G_x \le G(\Delta)$ on the sublevel set, and by Young's inequality $L_1 \sqrt{2 L_0 \Delta} = \sqrt{2 L_0 \cdot L_1^2 \Delta} \le L_0 + \frac{L_1^2 \Delta}{2}$, so $L_0 + L_1 G(\Delta) = L_0 + 2 L_1^2 \Delta + L_1 \sqrt{2 L_0 \Delta} \le 2 L_0 + \frac{5}{2} L_1^2 \Delta$. At $\Delta = \bar{\Delta} = L_0 / L_1^2$ this is $\frac{9}{2} L_0 \le 5 L_0$, and $G(\bar{\Delta}) = 2\bar{G} + \sqrt{2}\,\bar{G} \le \frac{7}{2}\bar{G}$.

□

A.1.4 Proof of Lemma 7

The claim is convexity-free: differentiability, Assumption 2 and $X^\star \neq \emptyset$ suffice. Since $\nabla f(x^\star) = 0$ by Fermat's condition, the fundamental theorem of calculus along $x_t^\star := x^\star + t(x^0 - x^\star)$ gives the identity $\Delta_{\text{init}} = \int_0^1 \langle \nabla f(x_t^\star),\, x^0 - x^\star \rangle \, dt$ (valid for the actual gap, not for an arbitrary certificate), and Lemma 6 with base $x^\star$ (where $G_{x^\star} = 0$) yields $G_{x_t^\star} \le \frac{L_0}{L_1}(e^{L_1 t R_0} - 1)$, hence

$$\begin{aligned}\Delta_{\text{init}} &= \int_0^1 \langle \nabla f(x_t^\star),\, x^0 - x^\star \rangle \, dt \\ &\le R_0 \int_0^1 \tfrac{L_0}{L_1}\big(e^{L_1 t R_0} - 1\big)\, dt \\ &= \tfrac{L_0}{L_1^2}\big(e^{L_1 R_0} - 1 - L_1 R_0\big) = \tfrac{L_0}{L_1^2}\,\psi(L_1 R_0).\end{aligned}$$

Finally $\psi(t) = \sum_{k\ge 2} \frac{t^k}{k!} = \frac{t^2}{2}\sum_{k\ge 0} \frac{2t^k}{(k+2)!} \le \frac{t^2}{2} e^t$ since $2/(k+2)! \le 1/k!$. Moreover $\psi(t)/t^2 = \sum_{k\ge 0} \frac{t^k}{(k+2)!}$ is increasing in $t \ge 0$ with limit $\frac{1}{2}$ at $t = 0$; hence Δ_{def} extends continuously to $\Delta_{\text{def}} = \frac{L_0 R_0^2}{2}$ at $L_1 = 0$ and is monotone in L_1, so the $L_{1,\text{eff}}$-substitution of Algorithm 1 only enlarges it.

□

A.1.5 Proof of Lemma 3

The problem (5) is convex with strongly feasible constraint (c is interior when $r < \infty$; when $r = \infty$ there is no constraint), so KKT conditions hold at $\hat{x}$: there is $\eta \ge 0$ (with $\eta = 0$ if $\|\hat{x} - c\| < r$) such that

$$\nabla f(\hat{x}) = -(\lambda + \eta)(\hat{x} - c). \tag{6}$$

If $\hat{x} = c$ then $\nabla f(c) = 0$ by (6) and the claim is trivial. Otherwise, monotonicity of ∇f gives $0 \le \langle \nabla f(\hat{x}) - \nabla f(c),\, \hat{x} - c\rangle = -(\lambda + \eta)\,\|\hat{x} - c\|^2 - \langle \nabla f(c),\, \hat{x} - c\rangle$, hence $(\lambda + \eta)\,\|\hat{x} - c\|^2 \le \|\nabla f(c)\|\,\|\hat{x} - c\|$ and therefore $\|\nabla f(\hat{x})\| = (\lambda + \eta)\,\|\hat{x} - c\| \le \|\nabla f(c)\|$. The same inequality gives $\|\hat{x} - c\| \le \|\nabla f(c)\|/(\lambda + \eta) \le \|\nabla f(c)\|/\lambda$. For $r = \infty$ one has $\eta = 0$, and (6) yields $\|\hat{x} - c\| = \|\nabla f(\hat{x})\|/\lambda$. The Fejér inequality is proved in the next paragraph.

□

A.1.6 Proof of the Fejér part of Lemma 3

Let $\lambda > 0$, let $\hat{x}$ be the minimizer of $f + \frac{\lambda}{2}\,\|\cdot - c\|^2$ over $\mathbb{B}(c,\, r)$ with $r \in (0, \infty]$, and let $u \in X^\star$, so that $\nabla f(u) = 0$. By the optimality conditions (6) there is $\eta \ge 0$ with $\nabla f(\hat{x}) = -(\lambda + \eta)(\hat{x} - c)$, and monotonicity of ∇f gives

$$0 \;\le\; \langle \nabla f(\hat{x}) - \nabla f(u),\, \hat{x} - u\rangle \;=\; -(\lambda + \eta)\,\langle \hat{x} - c,\, \hat{x} - u\rangle\,. \tag{7}$$

Since $\lambda + \eta \ge \lambda > 0$, (7) yields $\langle c - \hat{x},\, \hat{x} - u\rangle \ge 0$, whence

$$\|c - u\|^2 = \|c - \hat{x}\|^2 + 2\,\langle c - \hat{x},\, \hat{x} - u\rangle + \|\hat{x} - u\|^2 \;\ge\; \|c - \hat{x}\|^2 + \|\hat{x} - u\|^2\,. \tag{8}$$

This proves the lemma for $r = \infty$ and, with the same argument, the ball-constrained version used for the finisher and the strong-growth hops.

The restriction $\lambda > 0$ is necessary and is satisfied at every call made by our algorithms, since $\lambda_j \ge \Lambda_j > 0$ at every level (Appendix E.1) and $\lambda_{\text{cert}} > 0$ at the finisher. It cannot be dropped from the distance and Fejér conclusions: if $\lambda = 0$ and the ball constraint is inactive at some minimizer (e.g. f constant on a neighbourhood of c), then $\lambda + \eta = 0$, the direction of $\hat{x} - c$ is unconstrained by (7), the constrained minimizer need not be unique, and an arbitrary selection $\hat{x} \in \mathbb{B}(c,\, r)$ can violate (8). Only the gradient-norm conclusion $\|\nabla f(\hat{x})\| \le \|\nabla f(c)\|$ extends to $\lambda = 0$, by the same KKT and monotonicity argument; the displacement bound (which divides by λ), the Fejér inequality, and the unconstrained distance identity do not.

□

A.1.7 Proof of Lemma 4

Let $x, y \in \mathbb{B}(c,\, 2r)$ with $r \le \frac{1}{2L_1}$. By Lemma 6, $G_x \le e^{2L_1 r}(G_c + 2L_0 r) \le e\,(\hat{G} + L_0/L_1)$, the first claim. For the two-point bound, the segment $[x, y]$ lies in $\mathbb{B}(c,\, 2r)$ and has length $\|x - y\| \le 4r \le \frac{2}{L_1}$; split it into $m \le 2$ pieces of length at most $\frac{1}{L_1}$ with endpoints $u_0 = x, \ldots, u_m = y$. On each piece, (2) with base u_i and the first claim give $\|\nabla f(u_{i+1}) - \nabla f(u_i)\| \le \big(L_0 + L_1\, e(\hat{G} + L_0/L_1)\big)\, \|u_{i+1} - u_i\| = \big((1+e)L_0 + eL_1\hat{G}\big)\, \|u_{i+1} - u_i\|$. Summing over the pieces and using $1 + e \le 4$, $e \le 4$ yields $\|\nabla f(x) - \nabla f(y)\| \le 4(L_0 + L_1\hat{G})\, \|x - y\| = L_{\text{ball}}\, \|x - y\|$. The descent inequality along any segment $[x, y] \subset \mathbb{B}(c,\, 2r)$ follows by the integration argument of Lemma 1, using the two-point bound at the pairs $(x, x + t(y - x))$.

□

A.1.8 The Hessian form of generalized smoothness

Remark 4. *If f is twice differentiable with $\big\|\nabla^2 f(z)\big\| \le L_0 + L_1\, \|\nabla f(z)\|$ for all z, then Assumption 2 holds with the rescaled constants $(L_0', L_1') := (eL_0, eL_1)$. We must verify (2) for the pair (L_0', L_1'), i.e. on* its *admissible range $\|y - x\| \le 1/L_1' = 1/(eL_1)$. We do so by proving the required bound on the* larger *range $\|y - x\| \le 1/L_1$, which contains $[0, 1/(eL_1)]$. Fix x, y with $r := \|y - x\| \le 1/L_1$ and set $\varphi(t) := \|\nabla f(x + t(y - x))\|$; then $\varphi'(t) \le r\,(L_0 + L_1\varphi(t))$ for a.e. $t \in [0, 1]$, so Grönwall's inequality gives $L_0 + L_1\varphi(t) \le (L_0 + L_1 G_x)\, e^{L_1 r t}$, whence $\|\nabla f(y) - \nabla f(x)\| \le \int_0^1 \big\|\nabla^2 f(x + t(y - x))\big\|\, r\, dt \le r(L_0 + L_1 G_x)\frac{e^{L_1 r} - 1}{L_1 r} \le e\,(L_0 + L_1 G_x)\, r = (L_0' + L_1' G_x)\, r$, using $\frac{e^s - 1}{s} \le e$ for $s = L_1 r \le 1$. (For $L_1 = 0$ the Grönwall step divides by $L_1 r$; the conclusion then follows directly, $\varphi' \le L_0 r$ giving $\|\nabla f(y) - \nabla f(x)\| \le L_0 r$, or by continuity in L_1.) Since this bound holds for every $r \le 1/L_1$, it holds in particular for every $r \le 1/(eL_1)$, which is exactly the admissible range of the pair (L_0', L_1'); hence Assumption 2 holds with (L_0', L_1'). (The two ranges must be kept separate: the bound is* proved *on $r \le 1/L_1$ but only* needed*, and used, on the smaller rescaled range $r \le 1/(eL_1)$.) The factor e in the rescaling can be improved to $1/\ln 2$: what the argument needs is only $\frac{e^s - 1}{s} \le c$ for all $s \le \frac{1}{c}$, and since $\frac{e^s - 1}{s}$ is increasing the constraint binds at $s = \frac{1}{c}$, giving $e^{1/c} - 1 \le 1$; the optimal fixed point is $e^{1/c} = 2$, i.e. $c = \frac{1}{\ln 2}$. Hence the Hessian form $\big\|\nabla^2 f(z)\big\| \le L_0 + L_1\, \|\nabla f(z)\|$ implies Assumption 2 already with the constants $(L_0/\ln 2, L_1/\ln 2)$, the optimal factor for this argument.*

B Concentration toolbox

We record the probabilistic facts used throughout; all constants are absolute. A random vector $\zeta \in \mathbb{R}^d$ is *σ-norm-sub-Gaussian* if $\sigma > 0$ and $\mathbb{E}\exp(\|\zeta\|^2/\sigma^2) \le e$; at $\sigma = 0$ this terminology means $\zeta = 0$ almost surely, as stipulated in Assumption 3. For $\sigma > 0$, Jensen's inequality implies $\mathbb{E}\,\|\zeta\|^2 \le \sigma^2$, and Markov's inequality gives, for any $s > 0$,

$$\mathbb{P}\big[\,\|\zeta\|^2 \ge \sigma^2(1 + s)\big] \le e \cdot e^{-(1+s)} = e^{-s}. \tag{9}$$

Lemma 8 (Batch estimates; Jin et al., 2019)**.** *There are absolute constants $c_g \ge 1$, $c_b \ge 1$ such that if $\zeta_1, \ldots, \zeta_B$ are independent, zero-mean and σ-norm-sub-Gaussian conditionally on the past, then (i) for any $\beta \in (0, 1)$, with probability at least $1 - \beta$, $\big\|\frac{1}{B}\sum_i \zeta_i\big\| \le c_g\sigma\sqrt{\log(2/\beta)/B}$; and (ii) $\frac{1}{B}\sum_i \zeta_i$ is $(c_b\sigma/\sqrt{B})$-norm-sub-Gaussian.*

Lemma 9 (Azuma for norm-sub-Gaussian martingale difference sequence (MDS); Jin et al., 2019)**.** *There is an absolute constant $c_{\text{Az}} \ge 1$ such that if $(\zeta_t)_{t \le N}$ are martingale differences with ζ_t ς-norm-sub-Gaussian conditionally on $\mathcal{F}_{t-1}$, and $(w_t)_{t \le N}$ are $\mathcal{F}_{t-1}$-measurable vectors, then with probability at least $1 - \beta$, $\big|\sum_{t \le N} \langle \zeta_t,\, w_t \rangle\big| \le c_{\text{Az}}\,\varsigma\sqrt{\sum_{t \le N} \|w_t\|^2\, \log(2/\beta)}$ whenever $\|w_t\| \le W$ deterministically for all t (in which case $\sum \|w_t\|^2 \le NW^2$ may be used).*

Lemma 10 (Uniform bound on squared noise)**.** *If $(\zeta_t)_{t \le N}$ are each ς-norm-sub-Gaussian conditionally on the past, then with probability at least $1 - \beta$, $\max_{t \le N} \|\zeta_t\|^2 \le \varsigma^2\big(1 + \log(N/\beta)\big)$, and consequently for any deterministic weights a_t, $\sum_t a_t^2\, \|\zeta_t\|^2 \le \varsigma^2(1 + \log(N/\beta))\sum_t a_t^2$.*

Proof. Apply (9) with $s = \log(N/\beta)$ to each t and take a union bound. □

We use the explicit estimator GradEst(x, B) defined in (3). Under Assumption 3 and Lemma 8, it deviates from $\nabla f(x)$ by at most $c_g\sigma\sqrt{\log(2/\beta)/B}$ with probability $1-\beta$ and is $(c_b\sigma/\sqrt{B})$-norm-sub-Gaussian around it.

C THE INNER STOCHASTIC SOLVER: CONSTRAINED BATCHED STM WITH EPOCH RESTARTS

Let $Q \subseteq \mathbb{R}^d$ be a closed convex set, F convex and L-smooth on Q with minimizer $x_Q \in Q$, and let the oracle for F have variance at most ς^2 per call. We use STM exactly as stated in Procedure S1 of the main paper; in ARC-SG, $F = F_j$ is a proximal subproblem, Q is a ball, and batching reduces $\varsigma^2 = \sigma^2/b$. This section proves the high-probability bound of Proposition 1 and records the exact restart and batch schedule.

Projection model. Throughout the paper, the theory assumes an *exact* Euclidean projection oracle: STM performs one projection onto Q per iteration, and R-STM projects onto the two-ball intersections $Q_k = \mathbb{B}\big(y^{(k)},\, r_k\big) \cap Q$. Projection calls are counted separately from the stochastic-gradient oracle calls that our complexity bounds report—one projection per STM iteration, hence $\lceil 24\sqrt{L/\lambda}\,\rceil$ per R-STM epoch—and no projection-error analysis is provided. The two-ball projection does not reduce to a single-multiplier problem in general but is obtained from a two-dimensional dual system, solvable to machine precision (Remark 5); the experiments of Section I use single-ball projections only (Appendix I).

For λ-strongly convex F (as every F_j is), R-STM wraps STM into epochs: epoch k runs STM for $N_{\mathrm{ep}} = \lceil 24\sqrt{(L+\lambda)/\lambda}\,\rceil$ iterations on the ball $Q_k = \mathbb{B}\big(y^{(k)},\, r_k\big) \cap Q$ with the *shrinking certified radius* $r_k = 2\sqrt{2\bar{\Delta}_k/\lambda}$, where $\bar{\Delta}_k$ is the running high-probability bound on the F-gap, halved each epoch by choosing the per-iteration batch b_k so that the noise terms of Proposition 1 are at most $\bar{\Delta}_k/4$. Strong convexity certifies $x_Q \in Q_k$ from $\bar{\Delta}_k$; shrinking r_k is what makes the statistical cost of epoch k scale as $\widetilde{O}(\sigma^2/(\lambda\bar{\Delta}_k))$—the rate of the smooth strongly convex case—rather than as $\sigma^2 D^2/\bar{\Delta}_k^2$.

The restarted conclusion and oracle complexity are stated in Proposition 1; the remainder of this section proves that statement and records the explicit restart schedule.

C.1 CONSTRAINED BATCHED STM: PROOF OF PROPOSITION 1

Procedure S1 in the main paper states the complete update; we prove its high-probability guarantee here.

Recall the setting: Q closed convex, F convex on Q and L-smooth along segments of Q (two-point Lipschitz gradients plus the descent inequality, as provided by Lemma 4), $x_Q \in \arg\min_Q F$, $D \geq \max_{x\in Q}\|x - x_Q\|$, and the oracle returns $\hat{g}(x) = \nabla F(x) + \zeta$ with ζ conditionally ς-norm-sub-Gaussian. In Procedure S1, $a_t = \frac{t}{4L}$ and $A_t = \sum_{s\leq t} a_s = \frac{t(t+1)}{8L}$ satisfy

$$A_{t+1} = A_t + a_{t+1}, \qquad La_{t+1}^2 = \tfrac{(t+1)^2}{16L} \leq \tfrac{(t+1)(t+2)}{16L} = \tfrac{A_{t+1}}{2}. \tag{10}$$

All iterates lie in Q (convex combinations and projections of points of Q). Write $\hat{g}_{t+1} := \hat{g}(x_{t+1}) = \nabla F(x_{t+1}) + \zeta_{t+1}$.

One-step inequality. Since $y_{t+1} - x_{t+1} = \frac{a_{t+1}}{A_{t+1}}(z_{t+1} - z_t)$ and $[x_{t+1}, y_{t+1}] \subset Q$, the descent inequality gives

$$A_{t+1}F(y_{t+1}) \leq A_{t+1}F(x_{t+1}) + a_{t+1}\left\langle \nabla F(x_{t+1}),\, z_{t+1} - z_t\right\rangle + \tfrac{La_{t+1}^2}{2A_{t+1}}\left\|z_{t+1} - z_t\right\|^2. \tag{11}$$

The projection step means $z_{t+1} = \arg\min_{z\in Q}\{a_{t+1}\left\langle \hat{g}_{t+1},\, z\right\rangle + \frac{1}{2}\left\|z - z_t\right\|^2\}$, whose optimality condition against any $u \in Q$ reads

$$a_{t+1}\left\langle \hat{g}_{t+1},\, z_{t+1} - u\right\rangle \leq \tfrac{1}{2}\left\|u - z_t\right\|^2 - \tfrac{1}{2}\left\|u - z_{t+1}\right\|^2 - \tfrac{1}{2}\left\|z_{t+1} - z_t\right\|^2. \tag{12}$$

From the definition of x_{t+1}, $a_{t+1}(x_{t+1} - z_t) = A_t(y_t - x_{t+1})$, so by convexity

$$a_{t+1} \langle \nabla F(x_{t+1}),\, x_{t+1} - z_t \rangle = A_t \langle \nabla F(x_{t+1}),\, y_t - x_{t+1} \rangle \le A_t \big(F(y_t) - F(x_{t+1}) \big), \tag{13}$$

and $a_{t+1} \langle \nabla F(x_{t+1}),\, u - x_{t+1} \rangle \le a_{t+1}(F(u) - F(x_{t+1}))$. Decompose $\langle \nabla F(x_{t+1}),\, z_{t+1} - z_t \rangle = \langle \nabla F(x_{t+1}),\, z_{t+1} - u \rangle + \langle \nabla F(x_{t+1}),\, u - x_{t+1} \rangle + \langle \nabla F(x_{t+1}),\, x_{t+1} - z_t \rangle$ and $\langle \nabla F(x_{t+1}),\, z_{t+1} - u \rangle = \langle \hat{g}_{t+1},\, z_{t+1} - u \rangle - \langle \zeta_{t+1},\, z_{t+1} - u \rangle$; for the noise term use Young's inequality:

$$\begin{aligned} -a_{t+1} \langle \zeta_{t+1},\, z_{t+1} - u \rangle &= -a_{t+1} \langle \zeta_{t+1},\, z_t - u \rangle - a_{t+1} \langle \zeta_{t+1},\, z_{t+1} - z_t \rangle \\ &\le -a_{t+1} \langle \zeta_{t+1},\, z_t - u \rangle + a_{t+1}^2 \left\| \zeta_{t+1} \right\|^2 + \tfrac{1}{4} \left\| z_{t+1} - z_t \right\|^2 . \end{aligned} \tag{14}$$

Combining (11)–(14) and $\frac{L a_{t+1}^2}{2A_{t+1}} \le \frac{1}{4}$ from (10), the $\|z_{t+1} - z_t\|^2$ terms cancel and

$$\begin{aligned} A_{t+1} F(y_{t+1}) \le A_t F(y_t) + a_{t+1} F(u) + \tfrac{1}{2} \left\| u - z_t \right\|^2 - \tfrac{1}{2} \left\| u - z_{t+1} \right\|^2 \\ + a_{t+1}^2 \left\| \zeta_{t+1} \right\|^2 - a_{t+1} \langle \zeta_{t+1},\, z_t - u \rangle . \end{aligned}$$

Telescoping over $t = 0, \dots, N-1$ with $u = x_Q$ and $A_0 = 0$:

$$A_N \big(F(y_N) - F(x_Q) \big) \le \tfrac{1}{2} \left\| x_Q - z_0 \right\|^2 + \sum_{t=1}^{N} a_t^2 \left\| \zeta_t \right\|^2 + \Big| \sum_{t=1}^{N} a_t \langle \zeta_t,\, x_Q - z_{t-1} \rangle \Big| . \tag{15}$$

Concentration. The weights $w_t := a_t(x_Q - z_{t-1})$ are predictable with $\|w_t\| \le a_t D$, and $\sum_t a_t^2 = \frac{1}{16L^2} \sum_{t \le N} t^2 \le \frac{N(N+1)^2}{48L^2}$. By Lemma 9 (with failure probability $\beta/2$) and Lemma 10 (with $\beta/2$),

$$\Big| \sum_t a_t \langle \zeta_t,\, x_Q - z_{t-1} \rangle \Big| \le c_{\mathrm{Az}} \varsigma D \tfrac{(N+1)\sqrt{N}}{\sqrt{48}\, L} \sqrt{\log \tfrac{4}{\beta}},$$

$$\sum_t a_t^2 \left\| \zeta_t \right\|^2 \le \varsigma^2 \big(1 + \log \tfrac{2N}{\beta} \big) \tfrac{N(N+1)^2}{48L^2}.$$

simultaneously with probability at least $1 - \beta$. Dividing (15) by $A_N = \frac{N(N+1)}{8L}$ and using $\|x_Q - z_0\| \le D$, $1 + \log(2N/\beta) \le 2\log(4N/\beta)$:

$$F(y_N) - F(x_Q) \le \frac{4LD^2}{N(N+1)} + \frac{8 c_{\mathrm{Az}}}{\sqrt{48}} \cdot \frac{\varsigma D \sqrt{\log(4/\beta)}}{\sqrt{N}} + \frac{(N+1)\, \varsigma^2 \log(4N/\beta)}{3L}.$$

Proposition 1 follows with $c_1 := 2c_{\mathrm{Az}} \ge 8c_{\mathrm{Az}}/\sqrt{48}$ and $c_2 := \frac{1}{3}$.

□

C.2 R-STM: PROOF OF PROPOSITION 1

The epoch wrapper itself is defined in the main paper. Here we record its exact batch schedule and prove Proposition 1. Let F be λ-strongly convex on $\mathbb{B}(c, 2r)$ and L_F-smooth there, and define $x_F^\star := \arg\min_{x \in \mathbb{B}(c,\, r)} F(x)$ (in Phase II this is the unconstrained prox by containment; in the finisher it may lie on the boundary—both are covered). Given $y^{(0)} \in \mathbb{B}(c, r)$ with $F(y^{(0)}) - F(x_F^\star) \le H_0$, set $K := \lceil \log_2(H_0/\delta) \rceil_+$, $\bar{\Delta}_k := H_0\, 2^{-k}$, and for $k = 0, \dots, K-1$ run STM for

$$N_{\mathrm{ep}} := \lceil 24 \sqrt{L_F/\lambda} \rceil \quad \text{iterations on} \quad Q_k := \mathbb{B}\Big(y^{(k)},\, r_k \Big) \cap \mathbb{B}(c,\, r)\,, \qquad r_k := 2\sqrt{2\bar{\Delta}_k/\lambda},$$

with per-iteration batch size

$$b_k := \max \Big\{ 1,\ \Big\lceil C_b \frac{\sigma^2 \log\big(8K N_{\mathrm{ep}}/\beta\big)}{\sqrt{\lambda L_F}\ \bar{\Delta}_k} \Big\rceil \Big\}, \qquad C_b := 48 c_1^2 c_b^2 + 70 c_b^2 + 1, \tag{16}$$

and let $y^{(k+1)}$ be the STM output. We verify the invariant $F(y^{(k)}) - F(x_F^\star) \le \bar{\Delta}_k$ by induction; it holds at $k = 0$. Given the invariant, strong convexity yields $\big\| y^{(k)} - x_F^\star \big\| \le \sqrt{2\bar{\Delta}_k/\lambda} = r_k/2$, so $x_F^\star \in Q_k$; also $Q_k \subseteq \mathbb{B}(c, r)$, so F is L_F-smooth along segments of Q_k and $D_k :=$

$\max_{x\in Q_k}\|x-x_F^\star\| \le r_k + r_k/2 = \frac{3}{2}r_k = 3\sqrt{2\bar\Delta_k/\lambda}$. Applying Proposition 1 on Q_k with confidence β/K and per-query sub-Gaussian parameter $\varsigma_k = c_b\sigma/\sqrt{b_k}$ (Lemma 8(ii)):

$$\begin{aligned} F(y^{(k+1)}) - F(x_F^\star) \le \underbrace{\frac{8L_F\cdot 18\bar\Delta_k/\lambda}{N_{\mathrm{ep}}^2}}_{\le \bar\Delta_k/4 \text{ by } N_{\mathrm{ep}}^2 \ge 576L_F/\lambda} \\ + \underbrace{\frac{c_1\varsigma_k D_k\sqrt{\log(4K/\beta)}}{\sqrt{N_{\mathrm{ep}}}}}_{\le \bar\Delta_k/8 \text{ by (16)}} \\ + \underbrace{\frac{c_2\varsigma_k^2(N_{\mathrm{ep}}+1)\log(4KN_{\mathrm{ep}}/\beta)}{L_F}}_{\le \bar\Delta_k/8 \text{ by (16)} \le \frac{\bar\Delta_k}{2}}. \end{aligned}$$

where the two noise bounds follow by substituting $\varsigma_k^2 = c_b^2\sigma^2/b_k$, $D_k^2 = 18\bar\Delta_k/\lambda$, $N_{\mathrm{ep}} \ge 24\sqrt{L_F/\lambda}$ and the two summands of C_b respectively. A union bound over the K epochs completes the induction with probability $1-\beta$, giving $F(y^{(K)}) - F(x_F^\star) \le \delta$ and, by strong convexity, $\|y^{(K)} - x_F^\star\| \le \sqrt{2\delta/\lambda}$. The total number of oracle calls is

$$\begin{aligned} \sum_{k=0}^{K-1} N_{\mathrm{ep}} b_k &\le KN_{\mathrm{ep}} + 25\sqrt{\tfrac{L_F}{\lambda}}\, C_b \frac{\sigma^2\log(8KN_{\mathrm{ep}}/\beta)}{\sqrt{\lambda L_F}} \sum_k \frac{1}{\bar\Delta_k} \\ &\le 25K\sqrt{\tfrac{L_F}{\lambda}} + \frac{100C_b\sigma^2\log(8KN_{\mathrm{ep}}/\beta)}{\lambda\delta}. \end{aligned}$$

using $\lceil 24\sqrt{L_F/\lambda}\rceil \le 25\sqrt{L_F/\lambda}$ (since $L_F \ge \lambda$) and $\sum_k \bar\Delta_k^{-1} \le 2^{K+1}/H_0 \le 4/\delta$. All queries lie in $Q_k \subseteq \mathbb{B}(c, r) \subseteq \mathbb{B}(c, 2r)$. □

Remark 5 (Projections and initial gaps)**.** *Each Q_k is the intersection of two Euclidean balls with nonempty interior. Its exact Euclidean projection does* not *reduce to a single-multiplier problem in general: both spherical boundaries can be active simultaneously, and the KKT system then carries two multipliers $(\eta_1, \eta_2) \ge 0$,*

$$z - \Pi_{Q_k}(z) = \eta_1\big(\Pi_{Q_k}(z) - y^{(k)}\big) + \eta_2\big(\Pi_{Q_k}(z) - c\big), \tag{17}$$

so that the projection is obtained from a two-dimensional dual problem (which is smooth, concave and solvable to machine precision by any standard method) or, alternatively, by Dykstra's alternating algorithm, which converges for this pair of convex sets and does so linearly under standard regularity conditions (satisfied here, since the interiors intersect). Only when one of the two constraints is known to be inactive does the one-dimensional reduction apply. This distinction is purely computational: throughout the paper, "complexity" means the number of stochastic gradient evaluations*, and the arithmetic cost of projections (as well as of any line searches in the methods we compare against) is not counted in it; we make this convention explicit here because the two costs are conflated easily.*

In Phase II *the initial gap of level j is bounded by convexity as*

$$F_j(c_j) - F_j(\hat x_j) \le \langle \nabla F_j(c_j),\, c_j - \hat x_j\rangle = \langle \nabla f(c_j),\, c_j - \hat x_j\rangle \le \hat G_j \cdot \frac{\hat G_j}{\lambda_j}, \tag{18}$$

so $H_{0,j} := \hat G_j^2/\lambda_j$ is a valid certified input; in the finisher, $H_0 := \hat G \cdot r_{\mathrm{safe}} = \hat G/(2L_1)$ works by the same argument.

D Proof of Proposition 2

Procedure 1 and its interpretation are given in the main paper. Here we prove Proposition 2 and then record the default-certificate refinement used in Remark 1.

Corollary 2 (Default certificate)**.** *Let $q := L_1R_0$. This corollary is a statement about the default certificate: when $\Delta_{\mathrm{cert}} = \Delta_{\mathrm{def}} = \frac{L_0}{L_1^2}\psi(q)$ of Lemma 7 one has $\log_+ \frac{2L_1^2\Delta_{\mathrm{def}}}{L_0} = \log_+\big(2\psi(q)\big) \le 1 + q + \ln 2$ (since $\psi(q) \le e^q$), hence the logarithmic factor $\Lambda_{\lg}$ of (19) is $\Theta(1+q)$, and the bounds*

of Proposition 2 become

$$T_1 \le 1 + 7\,q\,(1 + q + \ln 2),$$
$$T_1 B_1 = \widetilde{O}\Big((1+q)^2 + (1+q)^4\,\tfrac{\sigma^2 L_1^2}{L_0^2}\Big),$$

i.e., in the regime $q \gtrsim 1$, T_1 is bounded by a $\Theta(q^2)$ quantity and $T_1 B_1 = \widetilde{O}\big(q^2 + q^4\sigma^2 L_1^2/L_0^2\big)$, with $\widetilde{O}$ hiding only absolute constants and polylogarithmic factors in $(q, 1/\alpha)$. Thus, under the default gap certificate, the Phase I *entry price is quadratic rather than linear in q; with an externally supplied certificate the corresponding form is $\widetilde{O}\big(1+q+(1+q)\sigma^2 L_1^2/L_0^2\big)$. Whether the quadratic dependence is sharp is open: on some example classes the actual iteration count is $\Theta(q)$, and sharpening the analysis (or proving a matching lower bound) is an open problem we do not address.*

D.1 Proof

Throughout $L_1 > 0$ and the parameters are

$$\Lambda_{\mathrm{lg}} := \big\lceil 2 + 2\log_+ \tfrac{2L_1^2\Delta_{\mathrm{cert}}}{L_0}\big\rceil, \qquad \varepsilon_1 := \frac{\bar G}{8\Lambda_{\mathrm{lg}}}, \qquad B_1 := \max\Big\{1,\ \Big\lceil \frac{c_g^2\sigma^2 \log \frac{8(T_1^{\max}+1)}{\alpha}}{\varepsilon_1^2}\Big\rceil\Big\},$$
$$T_1^{\max} := \max\Big\{1,\ \big\lceil 6\Lambda_{\mathrm{lg}} L_1 R_0\big\rceil,\ 1 + \big\lceil 7 L_1 R_0 \log_+ \tfrac{2L_1^2\Delta_{\mathrm{cert}}}{L_0}\big\rceil\Big\}. \tag{19}$$

By default $\Delta_{\mathrm{cert}} := \frac{L_0}{L_1^2}\psi(L_1 R_0)$, the certified bound Δ_{def} of Lemma 7, which makes the algorithm executable from $(x^0, R_0, L_0, L_1, \sigma, \alpha, \varepsilon)$ alone; any user-provided upper bound $\Delta_{\mathrm{user}} \ge f(x^0) - f^\star$ may be substituted, with the algorithm consuming $\min\{\Delta_{\mathrm{user}}, \Delta_{\mathrm{def}}\}$, which can only tighten the logarithms. Define the good event $\mathcal{E}_1 := \{\|\hat g_t - \nabla f(x_t)\| \le \varepsilon_1\ \forall t \le T_1^{\max}\}$; by Lemma 8(i) and a union bound, $\mathbb{P}[\mathcal{E}_1] \ge 1 - \alpha/4$. We work on $\mathcal{E}_1$. Note $\Lambda_{\mathrm{lg}} \ge 4$, so $\varepsilon_1 \le \bar G/32$.

Step executed ⇒ super-critical gradient. If iteration t executes a step then $\|\hat g_t\| > \frac{3}{4}G_{\mathrm{stop}} = 3\bar G$, hence

$$G_t \ge \|\hat g_t\| - \varepsilon_1 \ge 3\bar G - \tfrac{\bar G}{32} \ge \tfrac{23}{8}\bar G, \qquad \frac{\varepsilon_1}{\|\hat g_t\|} \le \frac{1}{24\Lambda_{\mathrm{lg}}} \le \frac{1}{96}, \qquad \frac{\varepsilon_1}{G_t} \le \frac{1}{23\Lambda_{\mathrm{lg}}} \le \frac{1}{92}, \tag{20}$$

and consequently $\frac{22}{23}G_t \le \|\hat g_t\| \le \frac{24}{23}G_t$. The step length is $\gamma_t\|\hat g_t\| = \frac{\|\hat g_t\|}{2(L_0 + L_1\|\hat g_t\|)} \le \frac{1}{2L_1}$, so Lemma 1 applies at base x_t.

(i) Per-step decrease. By Lemma 1 with $y = x_{t+1} = x_t - \gamma_t\hat g_t$,

$$f(x_{t+1}) \le f(x_t) - \gamma_t\,\langle \nabla f(x_t),\, \hat g_t\rangle + \tfrac{L_0 + L_1 G_t}{2}\gamma_t^2\,\|\hat g_t\|^2 .$$

Now $\langle \nabla f(x_t), \hat g_t\rangle \ge \|\hat g_t\|^2 - \varepsilon_1\|\hat g_t\| \ge \frac{23}{24}\|\hat g_t\|^2$ by (20), and $\frac{(L_0+L_1G_t)\gamma_t}{2} = \frac{L_0+L_1G_t}{4(L_0+L_1\|\hat g_t\|)} \le \frac{23}{88}$ since $\|\hat g_t\| \ge \frac{22}{23}G_t$. Hence the bracket is at least $\frac{23}{24} - \frac{23}{88} = \frac{23}{33}$, and using $\|\hat g_t\|^2 \ge (\frac{22}{23})^2 G_t^2$ and $\gamma_t \ge \frac{23}{24}\cdot\frac{1}{2(L_0+L_1G_t)}$,

$$f(x_t) - f(x_{t+1}) \;\ge\; \frac{23}{33}\Big(\frac{22}{23}\Big)^2\frac{23}{24}\cdot\frac{G_t^2}{2(L_0 + L_1 G_t)} \;\ge\; \frac{3}{10}\cdot\frac{G_t^2}{L_0 + L_1 G_t} \;>\; 0. \tag{21}$$

In particular f is non-increasing along the trajectory. In the super-critical regime (20) we have $L_0 \le \frac{8}{23}L_1 G_t$, so $L_0 + L_1 G_t \le \frac{31}{23}L_1 G_t$ and

$$f(x_t) - f(x_{t+1}) \ge \frac{3}{10}\cdot\frac{23}{31}\cdot\frac{G_t}{L_1} \ge \frac{2G_t}{9L_1}. \tag{22}$$

(iii) Distance control. Let $r_t := \|x_t - x^\star\|$. Expanding and using convexity ($\langle \nabla f(x_t), x_t - x^\star\rangle \ge \Delta_t$),

$$r_{t+1}^2 = r_t^2 - 2\gamma_t\,\langle \hat g_t,\, x_t - x^\star\rangle + \gamma_t^2\,\|\hat g_t\|^2 \le r_t^2 - 2\gamma_t\Delta_t + 2\gamma_t\varepsilon_1 r_t + \gamma_t\cdot\frac{\|\hat g_t\|^2}{2(L_0 + L_1\,\|\hat g_t\|)}.$$

The map $u \mapsto \frac{u^2}{2(L_0+L_1u)}$ is increasing, so with $\|\hat{g}_t\| \le \frac{24}{23}G_t$ and Lemma 2, $\frac{\|\hat{g}_t\|^2}{2(L_0+L_1\|\hat{g}_t\|)} \le (\frac{24}{23})^2\frac{G_t^2}{2(L_0+L_1G_t)} \le \frac{12}{11}\Delta_t$, whence

$$r_{t+1}^2 \le r_t^2 - \tfrac{10}{11}\gamma_t\Delta_t + 2\gamma_t\varepsilon_1 r_t \le r_t^2 + 2\gamma_t\varepsilon_1 r_t,$$
$$2\gamma_t\varepsilon_1 = \frac{\varepsilon_1}{L_0 + L_1\|\hat{g}_t\|} \le \frac{\varepsilon_1}{3L_1\bar{G}} = \frac{1}{24\Lambda_{\lg}L_1}. \tag{23}$$

Case $L_1R_0 \le \frac{1}{24}$ (degenerate). One has $\Delta_{\mathrm{cert}} \le \Delta_{\mathrm{def}} = \frac{L_0}{L_1^2}\psi(L_1R_0) \le \frac{L_0}{L_1^2}\psi(\frac{1}{24}) \le 0.00089\,\frac{L_0}{L_1^2}$, since $L_1R_0 \le \frac{1}{24}$ in this branch and ψ is increasing; hence, by Lemma 2 applied with $\Delta_{\mathrm{init}} \le \Delta_{\mathrm{cert}}$, $G_0 \le 2L_1\Delta_{\mathrm{cert}} + \sqrt{2L_0\Delta_{\mathrm{cert}}} \le (0.0018+0.0422)\bar{G} \le 0.05\,\bar{G}$, so $\|\hat{g}_0\| \le G_0+\varepsilon_1 \le 0.09\,\bar{G} < 3\bar{G}$: the algorithm returns $x^{\mathrm{I}} = x^0$ at $t=0$, and all claims hold trivially (with $\|\nabla f(x^{\mathrm{I}})\| \le 0.05\bar{G}$).

Case $L_1R_0 > \frac{1}{24}$. We prove $r_t \le \frac{3}{2}R_0$ for all $t \le T_1^{\max}$ by induction; the budget-exhausted iterate $x_{T_1^{\max}+1}$ is returned only off the good event $\mathcal{E}_1$ in Phase I and therefore needs no distance guarantee. Assume it up to t. Summing (23), $r_{t+1}^2 \le R_0^2 + T_1^{\max}\cdot\frac{1}{24\Lambda_{\lg}L_1}\cdot\frac{3}{2}R_0 = R_0^2 + \frac{T_1^{\max}R_0}{16\Lambda_{\lg}L_1}$. We bound the last term by $\frac{5}{4}R_0^2$ in each branch of the maximum in (19), using $\Lambda_{\lg}L_1R_0 \ge \frac{\Lambda_{\lg}}{24} \ge \frac{1}{6}$: if $T_1^{\max} = \lceil 6\Lambda_{\lg}L_1R_0\rceil \le 6\Lambda_{\lg}L_1R_0 + 1$ then $\frac{T_1^{\max}R_0}{16\Lambda_{\lg}L_1} \le \frac{3}{8}R_0^2 + \frac{R_0}{16\Lambda_{\lg}L_1} \le \frac{3}{8}R_0^2 + \frac{6}{16}R_0^2 \le \frac{3}{4}R_0^2$; if $T_1^{\max} = 1 + \lceil 7L_1R_0\log_+(\cdot)\rceil \le 2 + 7L_1R_0\log_+(\cdot)$ then, since $7\log_+(\cdot) \le \frac{7}{2}\Lambda_{\lg}$, $\frac{T_1^{\max}R_0}{16\Lambda_{\lg}L_1} \le \frac{7\log_+(\cdot)}{16\Lambda_{\lg}}R_0^2 + \frac{2R_0}{16\Lambda_{\lg}L_1} \le \frac{7}{32}R_0^2 + \frac{3}{4}R_0^2 \le \frac{31}{32}R_0^2$; and $T_1^{\max} = 1$ is trivial. In all cases $r_{t+1}^2 \le R_0^2 + \frac{5}{4}R_0^2 = \frac{9}{4}R_0^2$, closing the induction: $r_t \le \frac{3}{2}R_0$.

(iv) Iteration count and output. Let $\bar{\Delta}_{\mathrm{stop}} := \frac{L_0}{2L_1^2}$. If $\Delta_t \le \bar{\Delta}_{\mathrm{stop}}$ then by Lemma 2 $G_t \le 2L_1\bar{\Delta}_{\mathrm{stop}} + \sqrt{2L_0\bar{\Delta}_{\mathrm{stop}}} = 2\bar{G} < \frac{23}{8}\bar{G}$, contradicting (20); hence every executed step has $\Delta_t > \bar{\Delta}_{\mathrm{stop}}$. Combining (22) with $G_t \ge \Delta_t/r_t \ge \frac{2\Delta_t}{3R_0}$ (convexity), $\Delta_{t+1} \le \big(1 - \frac{4}{27L_1R_0}\big)\Delta_t$ (if $\frac{4}{27L_1R_0} \ge 1$ a single step already gives $\Delta_{t+1} \le 0$). Therefore after at most $\frac{27}{4}L_1R_0\ln\frac{\Delta_{\mathrm{cert}}}{\bar{\Delta}_{\mathrm{stop}}} \le 7L_1R_0\log_+\frac{2L_1^2\Delta_{\mathrm{cert}}}{L_0}$ executed steps the gap is below $\bar{\Delta}_{\mathrm{stop}}$, and the next check sees $\|\hat{g}_t\| \le G_t + \varepsilon_1 \le 2\bar{G} + \frac{\bar{G}}{32} < 3\bar{G}$ and returns. Thus $T_1 \le 1 + 7L_1R_0\log_+\frac{2L_1^2\Delta_{\mathrm{cert}}}{L_0} \le T_1^{\max}$, and the returned point satisfies $\|\nabla f(x^{\mathrm{I}})\| \le \|\hat{g}_{T_1}\| + \varepsilon_1 \le 3\bar{G} + \frac{\bar{G}}{32} \le \frac{25}{8}\bar{G} \le \frac{7}{8}G_{\mathrm{stop}}$, with $f(x^{\mathrm{I}}) \le f(x^0)$ by (21) and $\|x^{\mathrm{I}} - x^\star\| \le \frac{3}{2}R_0$. The oracle-call count is

$$\begin{aligned} T_1B_1 &\le \Big(2 + 7L_1R_0\log_+\tfrac{2L_1^2\Delta_{\mathrm{cert}}}{L_0}\Big) \\ &\quad\times\Big(1 + 64c_g^2\Lambda_{\lg}^2\tfrac{\sigma^2L_1^2}{L_0^2}\log\tfrac{8(T_1^{\max}+1)}{\alpha}\Big) \\ &= \widetilde{O}\Big(1 + L_1R_0 + (1+L_1R_0)\tfrac{\sigma^2L_1^2}{L_0^2}\Big). \end{aligned} \tag{24}$$

the additive floors remain because the deterministic iteration count is at least one and $B_1 \ge 1$.

□

E THE COMPLETE ARC-SG SCHEDULE AND Phase II ANALYSIS

The complete outer logic is given by Phase II, called from the sole ARC-SG wrapper, Algorithm 1. The schedules below supply its exact constants. The remainder of this section proves the invariants, travel–halving count, certified exits, and complexity claimed in Proposition 3 and Theorem 1.

E.1 THE FULL PARAMETER SCHEDULE OF Phase II AND THE FINISHER

Standing convention. Throughout the Phase II analysis, L_1 stands for the effective parameter $L_{1,\mathrm{eff}} := \max\{L_1,\, 1/(4\bar{R})\}$ introduced in Algorithm 1; equivalently, we assume without loss of generality that

$$L_1 \ge \tfrac{1}{4\bar{R}} \qquad (\text{equivalently } L_1\bar{R} \ge \tfrac{1}{4},\ \bar{\kappa} \ge \tfrac{5}{4}). \tag{25}$$

This is indeed no loss: Assumption 2 is monotone in L_1 (if it holds with (L_0, L_1) it holds with (L_0, L_1') for any $L_1' \geq L_1$, on the smaller range $\|y - x\| \leq 1/L_1'$), so when the problem's L_1 is smaller than $1/(4\bar{R})$ the algorithm and the analysis may be run with the enlarged effective parameter $L_{1,\text{eff}}$ in place of L_1—a change of the algorithm's input, not a redefinition of the problem; every $\bar{\kappa}$-factor changes by a factor at most $\frac{5}{4}$, which is absorbed in $\widetilde{O}$. The convention also covers $L_1 = 0$ (the classical smooth case); under the substitution the effective threshold of Phase I is $3\bar{G} = 12L_0\bar{R}$, and since $G_{x^0} \leq 2L_1\Delta_{\text{cert}} + \sqrt{2L_0\Delta_{\text{cert}}}$ (Lemma 2, applicable since $\Delta_{\text{init}} \leq \Delta_{\text{cert}}$), the condition $\Delta_{\text{cert}} \leq L_0\bar{R}^2$ gives $G_{x^0} \leq 2 \cdot \frac{1}{4\bar{R}} \cdot L_0\bar{R}^2 + \sqrt{2L_0^2\bar{R}^2} = (\frac{1}{2} + \sqrt{2})L_0\bar{R} < 2L_0\bar{R} = \frac{\bar{G}}{2}$, hence $\|\hat{g}_0\| \leq G_{x^0} + \varepsilon_1 < \frac{\bar{G}}{2} + \frac{\bar{G}}{32} < 3\bar{G}$: Phase I exits at $t = 0$ whenever $\Delta_{\text{cert}} \leq L_0\bar{R}^2$—in the floor regime $L_1 \leq 1/(4\bar{R})$ unconditionally, since then $\Delta_{\text{cert}} \leq \Delta_{\text{def}}(L_{1,\text{eff}}) = 16L_0\bar{R}^2\psi(\frac{1}{12}) \approx 0.057L_0\bar{R}^2 < L_0\bar{R}^2$ by the monotonicity of Δ_{def} in the smoothness parameter (Lemma 7)—so Theorem 1 degenerates to the smooth-case complexity as claimed in Remark 6; when $\Delta_{\text{cert}} > L_0\bar{R}^2$ the (cheap, σ-scaled) Phase I iterations simply run as usual and their count is dominated by the stated bounds. From Algorithm 1 on, every L_1 appearing in parameters, algorithms and analysis denotes the effective parameter $L_{1,\text{eff}}$ (so in particular the finisher's $\lambda_{\text{cert}} = L_1\varepsilon/\bar{R}$ of (28) is positive even at $L_1 = 0$, as Lemma 3 requires).

Trivial regime. If the certified initial gap already meets the target, i.e. $\Delta_{\text{cert}} \leq \varepsilon$, the algorithm returns x^0 and there is nothing to prove; *all statements below therefore assume* $\varepsilon < \Delta_{\text{cert}}$, and this test is the first line of Algorithm 1. Since Δ_{cert} is computed explicitly—as the default Δ_{def} of Lemma 7, or as $\min\{\Delta_{\text{user}}, \Delta_{\text{def}}\}$ when a user certificate is supplied—this is a finite, checkable condition. The counters below are nevertheless defined with positive parts, so that every schedule quantity is well defined and positive for *all* $\varepsilon > 0$.

Global quantities. With $\bar{R} = 3R_0$, $\bar{\kappa} = 1 + L_1\bar{R}$, $\bar{G} = L_0/L_1$, and $\log_+ a = \max\{1, \log a\}$:

$$
\begin{aligned}
&S_G := \max\left\{1, \left\lceil \log_2^+ \frac{10L_0\bar{R}}{L_1\varepsilon} \right\rceil\right\}, \qquad S_\Lambda := \max\left\{1, \left\lceil \log_2^+ \frac{128L_0\bar{R}^2}{\varepsilon} \right\rceil\right\},\\
&J^{\text{sch}} := 60\lceil\bar{\kappa}\rceil S_G(S_\Lambda + 4) + 16,\\
&K_{\text{cert}} := \left\lceil 2(1 + L_1\bar{R})(S_G + 4) \right\rceil + 1,\\
&J_{\text{tot}} := J^{\text{sch}} + K_{\text{cert}},\\
&\iota_{\text{abs}} := \frac{L_0}{8L_1J_{\text{tot}}}, \qquad \iota_{\text{flr}} := \frac{L_0}{32L_1S_\Lambda},
\end{aligned}
\tag{26}
$$

where $\log_2^+ a := \max\{0, \log_2 a\}$. Both counters are thus at least 1, so $J^{\text{sch}}, K_{\text{cert}}, J_{\text{tot}} \geq 1$ and the denominators in $\iota_{\text{abs}}, \iota_{\text{flr}}$ never vanish. In the nontrivial regime $\varepsilon < \Delta_{\text{cert}}$ the positive part of S_Λ is inactive whenever $\varepsilon \leq L_0\bar{R}^2$, and that of S_G whenever $\varepsilon \leq 10\bar{G}\bar{R}$; since $10\bar{G}\bar{R} \geq L_0\bar{R}^2$ iff $L_1\bar{R} \leq 10$, the single condition $\varepsilon \leq L_0\bar{R}^2$ (a fortiori whenever $\Delta_{\text{cert}} \leq L_0\bar{R}^2$, e.g. in the floor regime $L_1 \leq 1/(4\bar{R})$) makes both inactive in the moderate regime $L_1\bar{R} \leq 10$, and then the bounds proved below coincide with the ones obtained from the unclipped definitions. When a positive part *is* active the corresponding counter equals 1 and every bound in this appendix only improves; for S_G this case means $\varepsilon > 10\bar{G}\bar{R}$, hence $\hat{G}_1 = 4\bar{G} < \frac{\varepsilon}{2\bar{R}}$ and termination (a) fires at level 1, so the clipped schedule is never actually executed.

Level quantities. At level j with state $(c_j, \hat{G}_j, \Lambda_j)$ (initialized $\hat{G}_1 = 4\bar{G}$, $\Lambda_1 = 32L_0$):

$$
\begin{aligned}
&L_j := 4(L_0 + L_1\hat{G}_j), \qquad \lambda_j := \max\{4L_1\hat{G}_j,\ \Lambda_j\};\\
&r_j := \frac{2\hat{G}_j}{\lambda_j}\ \Big(\le \frac{1}{2L_1}\Big), \qquad L_{F,j} := L_j + \lambda_j;\\
&\textit{gradient-driven level} \text{ if } 4L_1\hat{G}_j > \Lambda_j;\\
&\textit{floor level} \text{ if } 4L_1\hat{G}_j \le \Lambda_j;\\
&s_j := \begin{cases} \min\left\{\frac{R_0}{4J_{\mathrm{tot}}},\ \frac{\iota_{\mathrm{abs}}}{L_j}\right\} & \text{(gradient-driven and finisher levels)}\\ \min\left\{\frac{R_0}{8S_\Lambda},\ \frac{\iota_{\mathrm{flr}}}{L_j}\right\} & \text{(floor levels)}\end{cases}\\
&\delta_j := \min\left\{\frac{\lambda_j s_j^2}{2},\ \frac{\varepsilon}{8}\right\};\\
&\hat{G}'_{j+1} := \hat{G}_j + L_j s_j, \qquad \tau_j := \frac{\hat{G}'_{j+1}}{16};\\
&B_j^{\mathrm{est}} := \max\left\{1,\ \left\lceil \frac{c_g^2\sigma^2}{\tau_j^2}\log\frac{16J_{\mathrm{tot}}}{\alpha}\right\rceil\right\}.
\end{aligned}
\tag{27}
$$

For every normal or finisher level, set $\beta_j := \alpha/(4J_{\mathrm{tot}})$. Each level runs R-STM on $(F_j, \mathbb{B}(c_j,\, r_j))$ with strong convexity λ_j, smoothness $L_{F,j}$, certified initial gap $H_{0,j} := \hat{G}_j^2/\lambda_j$ (Remark 5), target δ_j, and failure probability β_j; the estimate $\hat{g}_j := \mathsf{GradEst}(c_{j+1}, B_j^{\mathrm{est}})$ has accuracy τ_j with failure probability $\alpha/(4J_{\mathrm{tot}})$ (Lemma 8); the new cap is $\hat{G}_{j+1} := \min\{\hat{G}'_{j+1},\ \|\hat{g}_j\| + \tau_j\}$, and the level is a *halving* level if $\hat{G}_{j+1} \le \frac{3}{4}\hat{G}_j$ (equivalently $\|\hat{g}_j\| \le \frac{3}{4}\hat{G}_j - \tau_j$). At floor levels, if $\Lambda_j \le \varepsilon/(2\bar{R}^2)$ the algorithm returns c_{j+1} (termination (b)); otherwise $\Lambda_{j+1} := \Lambda_j/2$. The gradient termination (a) returns c_j when $\hat{G}_j \le \varepsilon/(2\bar{R})$, checked at the head of every level (including finisher levels).

The finisher. Entered when $\hat{G}_j < 32\iota_{\mathrm{abs}}$ or when $j = J^{\mathrm{sch}}$ (Procedure 3); runs at most K_{cert} *finisher levels*, each a ball-constrained proximal step with

$$
\begin{aligned}
&\lambda_{\mathrm{cert}} := \frac{L_1\varepsilon}{\bar{R}}, \qquad r_{\mathrm{safe}} := \frac{1}{2L_1}, \qquad \theta := \frac{1}{2L_1\bar{R}},\\
&\delta_j^{\mathrm{cert}} := \min\left\{\frac{\varepsilon}{64L_1\bar{R}},\ \frac{\lambda_{\mathrm{cert}}s_j^2}{2}\right\}, \qquad H_{0,\mathrm{cert}} := \frac{\hat{G}_j}{2L_1},
\end{aligned}
\tag{28}
$$

solved by R-STM for the ball-constrained minimizer (Appendix C.2 covers this case verbatim), with the same cap bookkeeping (27) and the same s_j-formula as gradient-driven levels.

Confidence budget (single accounting). To avoid any double reading, Table 1 lists every stochastic event family of the analysis, its budget, and the number of instances it is split over, together with the unused reserved quarter. Normal and finisher levels draw from the *same* counters: the finisher's R-STM runs and measurements are among the $J_{\mathrm{tot}} = J^{\mathrm{sch}} + K_{\mathrm{cert}}$ instances already counted. Its remaining certification checks are deterministic on the three stochastic good events and therefore spend none of the reserved fourth quarter. Thus the *good event* $\mathcal{E}$ has failure probability at most $3\alpha/4$ and, in particular, $\mathbb{P}[\mathcal{E}] \ge 1 - \alpha$.

E.2 Analysis of Phase II: invariants, counting, the finisher, and Theorem 1

Throughout this section the good event $\mathcal{E}$ of Appendix E.1 holds, the convention (25) is in force, and "level" means a normal (Phase II) or finisher level; there are at most J_{tot} of them by construction. We abbreviate $\iota := \iota_{\mathrm{abs}}$.

Event family	Budget	Instances	Per instance
Phase I gradient estimates	$\alpha/4$	$T_1^{\max}+1$	$\alpha/(4(T_1^{\max}+1))$
R-STM runs (all levels, incl. finisher)	$\alpha/4$	J_{tot}	$\alpha/(4J_{\mathrm{tot}})$
Cap measurements (all levels, incl. finisher)	$\alpha/4$	J_{tot}	$\alpha/(4J_{\mathrm{tot}})$
Finisher-only certification events	$\alpha/4$ reserved	deterministic on the shared good event	not used

Table 1: Confidence budget. Each stochastic row is a separately budgeted event family; no independence or disjointness between rows is needed, since the budgets are combined by a union bound. Within each stochastic row the budget is split uniformly over a deterministically bounded number of instances. The fourth quarter is reserved for finisher-only checks. In the present construction those checks are deterministic on the shared good event, so the random events actually consume at most $3\alpha/4$, leaving slack.

Logical structure of the proof. The argument has one entry point and two possible exits, and each arrow below is a separate statement proved in this appendix:

$$\underbrace{\text{Phase I}}_{\text{Prop. 2}} \longrightarrow \underbrace{\text{invariants (I1)–(I5)}}_{\text{Lem. 12, 13}} \longrightarrow \begin{cases} \underbrace{\text{normal levels}}_{\text{Lem. 14: } \textit{per excursion}} \longrightarrow \underbrace{\text{gradient or floor exit}}_{\text{Lem. 15}} \\ \underbrace{\text{finisher}}_{\text{Lem. 5: } \textit{unconditional}} \longrightarrow \underbrace{\varepsilon\text{-solution}}_{\text{Thm. 1}} \end{cases} \tag{29}$$

The distinction matters and we flag it wherever it arises: Lemma 14 bounds the number of *productive* levels *within one gradient-scale excursion* and does not by itself bound the number of excursions; unconditional termination comes from the hard schedule cap J^{sch} together with the finisher (Lemma 5), which is the branch that makes Theorem 1 an unconditional statement.

To make this guarantee checkable in one place—rather than distributed across Lemmas 12–5—we record it as a single statement; the constituent lemmas below supply its ingredients, and the subsequent cost accounting yields Theorem 1.

Proposition 4 (Termination and exit correctness of Phase II)**.** *On $\mathcal{E}$, Phase II—together with the finisher (Procedure 3)—halts after executing at most $J_{\mathrm{tot}} = J^{\mathrm{sch}} + K_{\mathrm{cert}}$ levels, and, whichever of its three exits fires, the returned point $\hat{x}$ satisfies $f(\hat{x}) - f^\star \le \varepsilon$.*

Proof. Termination is by construction: the outer loop runs $j = 1, \dots, J^{\mathrm{sch}}$ and, at the latest when $j = J^{\mathrm{sch}}$, the schedule-cap condition in Phase II hands control to the finisher, which itself executes at most K_{cert} ball-constrained levels and then returns (Procedure 3); hence at most J_{tot} levels execute in all. *Exit correctness:* the invariants that make the three exits applicable—the certified caps $\|\nabla f(c_j)\| \le \hat{G}_j$ and the distances $\|c_j - x^\star\| \le \bar{R}$—hold at every executed level by Lemma 12. Given them, the gradient exit (a) certifies $f(\hat{x}) - f^\star \le \frac{\varepsilon}{2}$ and the floor exit (b) certifies $f(\hat{x}) - f^\star \le \frac{5}{8}\varepsilon$ (Lemma 15), while the finisher returns $\hat{x}$ with $f(\hat{x}) - f^\star \le \frac{7}{16}\varepsilon$ (Lemma 5); in every case $f(\hat{x}) - f^\star \le \varepsilon$.

□

E.2.1 Invariants

Lemma 11 (Floor-level count)**.** *On $\mathcal{E}$, Λ_j is non-increasing, halves at every non-terminal floor level, and never drops below $\frac{\varepsilon}{4\bar{R}^2}$; hence at most S_Λ floor levels execute in the entire run. Moreover, after a floor level i one has $\Lambda \ge 2L_1\hat{G}_i$, so the cap cannot halve between consecutive floor levels without the floor binding again.*

Proof. Λ changes only at floor levels, where it halves unless the level terminates; a floor level with $\Lambda_j \le \varepsilon/(2\bar{R}^2)$ returns, so all halvings start from $\Lambda_j > \varepsilon/(2\bar{R}^2)$ and $\Lambda \ge \varepsilon/(4\bar{R}^2)$ throughout. The number of halvings from $\Lambda_1 = 32L_0$ down to $\varepsilon/(4\bar{R}^2)$ is at most $\lceil \log_2 \frac{128L_0\bar{R}^2}{\varepsilon} \rceil = S_\Lambda$. At a binding floor level i, $\Lambda_i \ge 4L_1\hat{G}_i$, so after halving $\Lambda \ge 2L_1\hat{G}_i$; if later $\hat{G}_j \le \hat{G}_i/2$ then $4L_1\hat{G}_j \le 2L_1\hat{G}_i \le \Lambda_j$ and the floor binds again.

□

Lemma 12 (Invariants). *On $\mathcal{E}$, for every executed level j:*

(I1) Certified caps: $\|\nabla f(c_j)\| \le \hat{G}_j$.

(I2) Distances: $\|c_j - x^\star\| \le \frac{15}{8} R_0 \le \bar{R}$.

(I3) Local regularity: *F_j is λ_j-strongly convex and $L_{F,j}$-smooth along segments of $\mathbb{B}(c_j, 2r_j)$, where $r_j \le \frac{1}{2L_1}$ at all levels, and $L_j \ge L_0 + L_1 G_x$ pointwise on the ball (Lemma 4 with (I1)).*

(I4) Containment (normal levels): *$\|\hat{x}_j - c_j\| \le G_{c_j}/\lambda_j \le r_j/2$, so $\hat{x}_j$ is the unconstrained minimizer of F_j and $H_{0,j} = \hat{G}_j^2/\lambda_j$ certifies the initial gap (Remark 5). R-STM succeeds: $F_j(c_{j+1}) - F_j(\hat{x}_j) \le \delta_j$, $\|c_{j+1} - \hat{x}_j\| \le \sqrt{2\delta_j/\lambda_j} \le s_j$, and $c_{j+1} \in \mathbb{B}(c_j, r_j)$. At finisher levels the same holds with $\hat{x}_j$ replaced by the ball-constrained minimizer $\tilde{x}_j$.*

(I5) Cap-update validity: *$G_{c_{j+1}} \le \min\{\hat{G}_j + L_j s_j,\ \|\hat{g}_j\| + \tau_j\} = \hat{G}_{j+1}$, and each level raises the cap by at most $L_j s_j$ ($\le \iota$ at gradient-driven and finisher levels, $\le \iota_{\mathrm{flr}}$ at floor levels).*

Proof. Induction on j. *Base:* $\|\nabla f(c_1)\| \le \frac{25}{8}\bar{G} \le 4\bar{G} = \hat{G}_1$ and $\|c_1 - x^\star\| \le \frac{3}{2}R_0$ by Proposition 2. *Step:* given (I1), Lemma 4 with center c_j, cap $\hat{G}_j$ and radius r_j yields (I3); here $r_j = \frac{1}{2L_1}$ exactly at gradient-driven and finisher levels, and $r_j = 2\hat{G}_j/\Lambda_j \le \frac{1}{2L_1}$ at floor levels since $\Lambda_j \ge 4L_1\hat{G}_j$. For (I4), Lemma 3 with $r = \infty$ gives $\|\hat{x}_j - c_j\| \le G_{c_j}/\lambda_j \le \hat{G}_j/\lambda_j = r_j/2$, so the ball constraint is inactive and Proposition 1 applies with $x_Q = \hat{x}_j$ (assumptions hold by (I3)–(I4)); on $\mathcal{E}$ its conclusions hold, and $\sqrt{2\delta_j/\lambda_j} \le s_j$ because $\delta_j \le \lambda_j s_j^2/2$. For (I5): $\|\nabla f(\hat{x}_j)\| \le G_{c_j} \le \hat{G}_j$ by Lemma 3 (which covers the ball-constrained finisher prox), and the segment $[\hat{x}_j, c_{j+1}]$ lies in $\mathbb{B}(c_j, r_j)$, so by (I3), $G_{c_{j+1}} \le \|\nabla f(\hat{x}_j)\| + L_j\|c_{j+1} - \hat{x}_j\| \le \hat{G}_j + L_j s_j = \hat{G}'_{j+1}$; the anchor branch satisfies $\|\hat{g}_j\| + \tau_j \ge G_{c_{j+1}}$ on $\mathcal{E}$. Hence $\hat{G}_{j+1} \ge G_{c_{j+1}}$, proving (I1) at $j{+}1$, and $\hat{G}_{j+1} \le \hat{G}'_{j+1}$ bounds the rise. Finally (I2): by Lemma 3 (constrained version at finisher levels), $\|\hat{x}_j - x^\star\| \le \|c_j - x^\star\|$, so $\|c_{j+1} - x^\star\| \le \|c_j - x^\star\| + s_j$, and the s-budget of (27) gives $\sum_j s_j \le J_{\mathrm{tot}} \cdot \frac{R_0}{4J_{\mathrm{tot}}} + S_\Lambda \cdot \frac{R_0}{8S_\Lambda} = \frac{3R_0}{8}$ (at most S_Λ floor levels execute, by Lemma 11), whence $\|c_j - x^\star\| \le \frac{3}{2}R_0 + \frac{3}{8}R_0 = \frac{15}{8}R_0$.

□

Lemma 13 (Drift: cumulative rises and the cap ceiling). *On $\mathcal{E}$:*

$$\sum_j (\hat{G}_{j+1} - \hat{G}_j)_+ \le J_{\mathrm{tot}}\,\iota + S_\Lambda\,\iota_{\mathrm{flr}} \le \tfrac{L_0}{8L_1} + \tfrac{L_0}{32L_1} \le \tfrac{L_0}{6L_1}; \tag{30}$$

hence $\hat{G}_j \le \hat{G}_1 + \frac{L_0}{6L_1} \le 5\bar{G} =: \hat{G}_{\max}$ and $L_j \le 24L_0$ whenever $\hat{G}_j \le 5\bar{G}$; at floor levels additionally $\hat{G}_j \le \Lambda_j/(4L_1) \le 8\bar{G}$ and $L_j \le 36L_0$.

Proof. Immediate from Lemma 12(I5) and the level-type counts (at most J_{tot} levels total, at most S_Λ of them floor levels, by Lemma 11). □

E.2.2 Level counting

Lemma 14 (Level count). *On $\mathcal{E}$, during the normal phase (every executed gradient-driven level has $\hat{G}_j \ge 32\iota$ and $\hat{G}_j > \frac{\varepsilon}{2\bar{R}}$, else the finisher or termination (a) fires first):*

(i) Dichotomy and productive decrease: *at every non-halving gradient-driven level, $\|\nabla f(c_{j+1})\| \ge \frac{159}{256}\hat{G}_j$, hence $\|\nabla f(\hat{x}_j)\| \ge \frac{151}{256}\hat{G}_j$ and*

$$f(c_j) - f(c_{j+1}) \ \ge\ \frac{\|\nabla f(\hat{x}_j)\|^2}{2\lambda_j} - s_j\,\|\nabla f(\hat{x}_j)\| - \delta_j \ \ge\ \frac{\hat{G}_j}{29L_1}. \tag{31}$$

(ii) Per-excursion budget: *partition the reachable cap range* $(\frac{\varepsilon}{2\bar{R}}, \hat{G}_{\max}]$ *into the anchored dyadic bands* $\left[2^{t-1}\frac{\varepsilon}{2\bar{R}},\, 2^{t}\frac{\varepsilon}{2\bar{R}}\right)$, $t = 1, \dots, S_G$ *(these cover it by Lemma 13 and the definition of* S_G*). Within any maximal run of consecutive levels whose caps lie in one band* $[g, 2g)$*, the number* n *of non-halving gradient-driven levels obeys* $n \le 57L_1\bar{R}\,(2 + S_\Lambda/2) + h \le 114L_1\bar{R} + 29L_1\bar{R}S_\Lambda + h$*, where* h *is the number of halving levels in the same run.*

(iii) Halvings: *among any* J *executed normal levels, the number* H *of halving levels satisfies* $H \le 3S_G + \frac{J}{9} + 3S_\Lambda(S_G + 1)$.

Consequently, if *every dyadic band hosts at most* V *excursions, the normal phase terminates within* $60\lceil\bar{\kappa}\rceil S_G(S_\Lambda + 4)V + 16$ *levels; the potential argument does not by itself bound* V*, and the* unconditional *guarantee of the algorithm is supplied by the hard schedule cap* J^{sch} *(the* $V = 1$ *value of the above bound) together with the finisher, which preserves all guarantees whenever it is entered (Lemma 5).*

Proof. (i) Non-halving means $\|\hat{g}_j\| > \frac{3}{4}\hat{G}_j - \tau_j$; on $\mathcal{E}$, $\|\nabla f(c_{j+1})\| \ge \|\hat{g}_j\| - \tau_j > \frac{3}{4}\hat{G}_j - 2\tau_j$. Since $\hat{G}_j \ge 32\iota \ge 32L_j s_j$, $\tau_j = \frac{\hat{G}_j + L_j s_j}{16} \le \frac{33}{512}\hat{G}_j$, so $\|\nabla f(c_{j+1})\| \ge (\frac{3}{4} - \frac{33}{256})\hat{G}_j = \frac{159}{256}\hat{G}_j$, and by (I3)–(I4),

$$\|\nabla f(\hat{x}_j)\| \ge \|\nabla f(c_{j+1})\| - L_j\,\|c_{j+1} - \hat{x}_j\| \ge \tfrac{159}{256}\hat{G}_j - L_j s_j \ge (\tfrac{159}{256} - \tfrac{1}{32})\hat{G}_j = \tfrac{151}{256}\hat{G}_j. \quad (32)$$

We isolate the *transfer inequality*, which will be reused for the finisher: at every level (gradient-driven, floor, or finisher),

$$f(c_{j+1}) \;\le\; f(\hat{x}_j) \;+\; \delta_j \;+\; s_j\,\|\nabla f(\hat{x}_j)\|\,. \quad (33)$$

Proof of (33). Write $a := \|\hat{x}_j - c_j\|$ and $d := \|c_{j+1} - \hat{x}_j\|$, so $d \le s_j$ by (I4) and $\|c_{j+1} - c_j\| \ge (a - d)_+$ by the triangle inequality. From $F_j(c_{j+1}) \le F_j(\hat{x}_j) + \delta_j$,

$$f(c_{j+1}) \;=\; F_j(c_{j+1}) - \tfrac{\lambda_j}{2}\,\|c_{j+1} - c_j\|^2 \;\le\; f(\hat{x}_j) + \delta_j + \tfrac{\lambda_j}{2}\left[a^2 - (a-d)_+^2\right].$$

If $d \ge a$ then $a^2 - (a-d)_+^2 = a^2 \le ad$; if $d < a$ then $a^2 - (a-d)^2 = 2ad - d^2 \le 2ad$. In both cases $\frac{\lambda_j}{2}[a^2 - (a-d)_+^2] \le \lambda_j a\, d$. At normal levels containment makes the prox unconstrained, so first-order optimality gives $\lambda_j a = \|\nabla f(\hat{x}_j)\|$ exactly; at finisher levels the KKT conditions (6) give $(\lambda_j + \eta)a = \|\nabla f(\hat{x}_j)\|$ with $\eta \ge 0$, so $\lambda_j a \le \|\nabla f(\hat{x}_j)\|$ in all cases. Hence $\lambda_j a\, d \le d\,\|\nabla f(\hat{x}_j)\| \le s_j\,\|\nabla f(\hat{x}_j)\|$, proving (33). ◊

For the decrease in (31): $F_j(\hat{x}_j) \le F_j(c_j)$ gives $f(\hat{x}_j) \le f(c_j) - \frac{\lambda_j}{2}\,\|\hat{x}_j - c_j\|^2 = f(c_j) - \frac{\|\nabla f(\hat{x}_j)\|^2}{2\lambda_j}$ (unconstrained prox at gradient-driven levels), and combining with (33) yields the first inequality of (31). Numerically, at gradient-driven levels $\lambda_j = 4L_1\hat{G}_j$, so $\frac{\|\nabla f(\hat{x}_j)\|^2}{2\lambda_j} \ge \left(\frac{151}{256}\right)^2 \frac{\hat{G}_j}{8L_1} = \frac{22801}{524288}\cdot\frac{\hat{G}_j}{L_1}$, while

$$s_j\,\|\nabla f(\hat{x}_j)\| \le s_j\hat{G}_j \le \tfrac{\iota}{L_j}\hat{G}_j \le \tfrac{\iota}{4L_1} \le \tfrac{\hat{G}_j}{128L_1} = \tfrac{4096}{524288}\cdot\tfrac{\hat{G}_j}{L_1} \quad (34)$$

(using $L_j \ge 4L_1\hat{G}_j$ and $\iota \le \hat{G}_j/32$) and $\delta_j \le \frac{\lambda_j s_j^2}{2} \le \frac{\iota^2}{8L_1\hat{G}_j} \le \frac{\hat{G}_j}{8192L_1} = \frac{64}{524288}\cdot\frac{\hat{G}_j}{L_1}$; hence the net decrease is at least $\frac{22801-4096-64}{524288}\cdot\frac{\hat{G}_j}{L_1} = \frac{18641}{524288}\cdot\frac{\hat{G}_j}{L_1} \ge \frac{\hat{G}_j}{29L_1}$.

(ii) Fix an excursion through $[g, 2g)$ with n non-halving gradient-driven levels, h halving levels, and $\phi \le S_\Lambda$ floor levels (Lemma 11). Every in-band level can increase f by at most $\delta_j + s_j\,\|\nabla f(\hat{x}_j)\|$ (by (33) and $f(\hat{x}_j) \le f(c_j)$): at gradient-driven levels this is $\le \frac{\hat{G}_j}{120L_1} \le \frac{g}{60L_1}$ by the bounds above; at floor levels, $s_j\,\|\nabla f(\hat{x}_j)\| \le \frac{\iota_{\mathrm{flr}}}{L_j}\hat{G}_j \le \frac{\hat{G}_j}{128L_1S_\Lambda} \le \frac{g}{64L_1}$ (the middle step uses $L_j = 4(L_0 + L_1\hat{G}_j) \ge 4L_0$, so $\iota_{\mathrm{flr}}/L_j \le 1/(128L_1S_\Lambda)$) and $\delta_j \le \frac{\varepsilon}{8} \le \frac{g\bar{R}}{4}$ (as $g \ge \frac{\varepsilon}{2\bar{R}}$ by the band anchoring of (ii)), so the increase is at most $\frac{g\bar{R}}{2}$ (using $L_1\bar{R} \ge \frac{1}{4}$ from (25)). The certified gap at the first level of the excursion is $f(c) - f^\star \le G_c\,\|c - x^\star\| \le \hat{G}\bar{R} \le 2g\bar{R}$. Summing (31) over the n productive levels against this

budget, $n \cdot \frac{g}{29L_1} \le 2g\bar{R} + (n+h)\frac{g}{60L_1} + S_\Lambda \frac{g\bar{R}}{2}$, so $n(\frac{1}{29} - \frac{1}{60}) \le L_1\bar{R}(2 + \frac{S_\Lambda}{2}) + \frac{h}{60}$ and, since $\frac{1}{29} - \frac{1}{60} = \frac{31}{1740} \ge \frac{1}{57}$, $n \le 57L_1\bar{R}(2 + \frac{S_\Lambda}{2}) + h$.

(iii) Let $\Phi_j := \log_2\left(\hat{G}_j \cdot \frac{2\bar{R}}{\varepsilon}\right) \in (0, \log_2 \frac{10L_0\bar{R}}{L_1\varepsilon}] \subseteq (0, S_G]$ while the run continues (positivity because termination (a) has not fired, the upper bound by Lemma 13). A halving level decreases Φ by at least $\log_2 \frac{4}{3} \ge 0.41$; a non-halving gradient-driven or finisher level increases it by at most $\log_2(1 + \frac{\iota}{\hat{G}_j}) \le \log_2 \frac{33}{32} \le 0.0444 \le \frac{1}{9} \cdot 0.41$; a floor level increases it by at most $\log_2(1 + \frac{\iota_{\mathrm{flr}}}{\varepsilon/(2\bar{R})}) \le \log_2\left(1 + \frac{2^{S_G}}{160 S_\Lambda}\right) \le S_G + 1$. Telescoping Φ over J levels, using $\Phi \ge 0$ throughout and $\Phi_1 \le S_G$: $0.41\,H \le S_G + J \cdot 0.0444 + S_\Lambda(S_G + 1)$, i.e. $H \le 3S_G + \frac{J}{9} + 3S_\Lambda(S_G + 1)$.

For the final claim: with at most V excursions per band and S_G bands, the total is $J \le VS_G\big(114L_1\bar{R} + 29L_1\bar{R}S_\Lambda\big) + 2H + S_\Lambda$, and substituting (iii) and solving for J gives $J \le \frac{9}{7}\big(VS_G\bar{\kappa}(114 + 29S_\Lambda) + 6S_G + 6S_\Lambda(S_G + 1) + S_\Lambda\big) \le 60\lceil\bar{\kappa}\rceil S_G(S_\Lambda + 4)V + 16$, which equals J^{sch} for $V = 1$; the unconditional statement is the hard cap plus Lemma 5.

□

E.2.3 Correctness of the terminations

Lemma 15 (Exits (a) and (b))**.** *On $\mathcal{E}$: if termination (a) fires, $f(c_j) - f^\star \le \hat{G}_j\bar{R} \le \frac{\varepsilon}{2}$. If termination (b) fires at a floor level, then $f(c_{j+1}) - f^\star \le \frac{\Lambda_j}{2}\bar{R}^2 + \frac{\hat{G}_j^2}{2\Lambda_j} + \delta_j \le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} \le \frac{5}{8}\varepsilon$.*

Proof. (a) is Lemma 12(I1)–(I2) and convexity: $f(c_j) - f^\star \le \langle \nabla f(c_j), c_j - x^\star\rangle \le \hat{G}_j\bar{R}$. For (b): by containment $\hat{x}_j$ is the unconstrained minimizer of F_j, so $F_j(\hat{x}_j) \le F_j(x^\star)$ gives $f(\hat{x}_j) - f^\star \le \frac{\Lambda_j}{2}\|c_j - x^\star\|^2 \le \frac{\Lambda_j}{2}\bar{R}^2 \le \frac{\varepsilon}{4}$; and

$$f(c_{j+1}) \le F_j(c_{j+1}) \le F_j(\hat{x}_j) + \delta_j = f(\hat{x}_j) + \tfrac{\Lambda_j}{2}\|\hat{x}_j - c_j\|^2 + \delta_j \le f(\hat{x}_j) + \tfrac{\hat{G}_j^2}{2\Lambda_j} + \delta_j, \quad (35)$$

where $\frac{\hat{G}_j^2}{2\Lambda_j} \le \frac{\Lambda_j}{32L_1^2} \le \frac{\varepsilon}{64L_1^2\bar{R}^2} \le \frac{\varepsilon}{4}$ using $\hat{G}_j \le \Lambda_j/(4L_1)$, $\Lambda_j \le \frac{\varepsilon}{2\bar{R}^2}$ and (25).

□

E.2.4 The finisher

Procedure 3 in the main paper states the complete ball-constrained certification routine.

Proof of Lemma 5. Let $\mathrm{gap}_k := f(c^{(k)}) - f^\star$ along the finisher iterates and write $\lambda := \lambda_{\mathrm{cert}}$, $\tilde{x}_k$ for the exact constrained minimizer of $F_k := f + \frac{\lambda}{2}\|\cdot - c^{(k)}\|^2$ over $\mathbb{B}\big(c^{(k)}, r_{\mathrm{safe}}\big)$. Note $\theta = \frac{1}{2L_1\bar{R}} \le 2$ by (25).

Per-level transfer. By the two-case argument in the proof of (33), with $a := \|\tilde{x}_k - c^{(k)}\| \le r_{\mathrm{safe}}$ and $d := \|c^{(k+1)} - \tilde{x}_k\| \le s$,

$$\begin{aligned} f(c^{(k+1)}) &\le f(\tilde{x}_k) + \delta^{\mathrm{cert}} + \lambda\, a\, d \\ &\le f(\tilde{x}_k) + \delta^{\mathrm{cert}} + \lambda\, r_{\mathrm{safe}}\, s \\ &\le f(\tilde{x}_k) + \tfrac{\theta\varepsilon}{32} + \tfrac{\varepsilon s}{2\bar{R}} \\ &\le f(\tilde{x}_k) + \tfrac{\theta\varepsilon}{16}, \end{aligned}$$

using $\delta^{\mathrm{cert}} \le \frac{\varepsilon}{64L_1\bar{R}} = \frac{\theta\varepsilon}{32}$, $\lambda r_{\mathrm{safe}} = \frac{\varepsilon}{2\bar{R}}$, and $\frac{\varepsilon s}{2\bar{R}} \le \frac{\varepsilon R_0}{8J_{\mathrm{tot}}\bar{R}} \le \frac{\theta\varepsilon}{32}$ since $s \le \frac{R_0}{4J_{\mathrm{tot}}}$ and $J_{\mathrm{tot}} \ge \frac{8L_1\bar{R}}{3}$.

Case $\theta \ge 1$ (i.e. $2L_1\bar{R} \le 1$). Then $r_{\mathrm{safe}} = \frac{1}{2L_1} \ge \bar{R} \ge \|c^{(k)} - x^\star\|$, so $x^\star$ is feasible at every level and $F_k(\tilde{x}_k) \le F_k(x^\star)$ gives $\mathrm{gap}(\tilde{x}_k) \le \frac{\lambda}{2}\|c^{(k)} - x^\star\|^2 \le \frac{L_1\varepsilon\bar{R}}{2} \le \frac{\varepsilon}{4}$; with the transfer ($\theta \le 2$), every level independently returns $\mathrm{gap}_{k+1} \le \frac{\varepsilon}{4} + \frac{2\varepsilon}{16} = \frac{3\varepsilon}{8} \le \frac{7\varepsilon}{16}$.

Case $\theta < 1$. If $\|c^{(k)} - x^\star\| \le r_{\text{safe}}$ then, as above, $\text{gap}(\tilde{x}_k) \le \frac{\lambda}{2} r_{\text{safe}}^2 = \frac{\varepsilon}{8L_1\bar{R}} = \frac{\theta\varepsilon}{4}$. Otherwise put $u := c^{(k)} + \theta'(x^\star - c^{(k)})$ with $\theta' := r_{\text{safe}} / \|c^{(k)} - x^\star\| \in [\theta, 1)$ (Lemma 12(I2) gives $\|c^{(k)} - x^\star\| \le \bar{R}$, maintained through finisher levels by the constrained Fejér property, Lemma 3, and the s-budget); then $u \in \mathbb{B}(c^{(k)}, r_{\text{safe}})$ and by convexity $\text{gap}(\tilde{x}_k) \le F_k(u) - f^\star \le (1-\theta')\,\text{gap}_k + \frac{\lambda}{2} r_{\text{safe}}^2 \le (1-\theta)\,\text{gap}_k + \frac{\theta\varepsilon}{4}$. Combining with the transfer, in both branches $\text{gap}_{k+1} \le (1-\theta)\,\text{gap}_k + \frac{\theta\varepsilon}{4} + \frac{\theta\varepsilon}{16} = (1-\theta)\,\text{gap}_k + \frac{5\theta\varepsilon}{16}$, and unrolling, $\text{gap}_k \le (1-\theta)^k\,\text{gap}_0 + \frac{5\varepsilon}{16}$. The entry gap obeys $\text{gap}_0 \le \hat{G}\bar{R} \le 5\bar{G}\bar{R}$ (Lemmas 12, 13), so after $k^* := \lceil \frac{1}{\theta} \ln \frac{40L_0\bar{R}}{L_1\varepsilon} \rceil \le 2L_1\bar{R} \cdot 0.7(S_G + 2) + 1 \le K_{\text{cert}}$ levels, $(1-\theta)^{k^*}\,\text{gap}_0 \le \frac{\varepsilon}{8}$ and $\text{gap}_{K_{\text{cert}}} \le \frac{\varepsilon}{8} + \frac{5}{16}\varepsilon = \frac{7}{16}\varepsilon$. Cap bookkeeping remains valid at finisher levels by Lemma 12 (whose proof covers them), so an early exit via (a) is also correct. □

E.2.5 Cost accounting and the full theorem

Umbrella logarithms. At level j, R-STM runs $K_j = \lceil \log_2(H_{0,j}/\delta_j) \rceil_+$ epochs of $N_{\text{ep},j} = \lceil 24\sqrt{L_{F,j}/\lambda_j} \rceil$ iterations. Using the lower bounds on $\lambda_j s_j$ established below and $H_{0,j} \le \hat{G}_j^2/\lambda_j$ (resp. $\hat{G}_j/(2L_1)$ at finisher levels), a routine check gives $K_j \le \bar{K} := 20 + 2\lceil \log_2 J_{\text{tot}} \rceil + \lceil \log_2 \big(1 + \frac{96L_0\bar{R}^2}{\varepsilon}\big) \rceil$ at every level, and every logarithm appearing in Proposition 1 at any level is bounded by $\ell := \log\big(2^{20} J_{\text{tot}}^3 \big(1 + \frac{96L_0\bar{R}^2}{\varepsilon}\big)^2/\alpha\big)$. By Proposition 1 (confidence $\alpha/(4J_{\text{tot}})$), level j costs at most $25\bar{K}\sqrt{1 + L_j/\lambda_j} \;+\; 96C_b\sigma^2\ell/(\lambda_j\delta_j) \;+\; B_j^{\text{est}}$ oracle calls.

Deterministic-equivalent part. By Lemma 13 and $\hat{G}_j > \frac{\varepsilon}{2\bar{R}}$: at gradient-driven levels $\sqrt{1 + L_j/\lambda_j} = \sqrt{2 + \bar{G}/\hat{G}_j} \le 2 + \sqrt{2L_0\bar{R}/(L_1\varepsilon)}$; at floor levels $\sqrt{1 + 36L_0 \cdot 4\bar{R}^2/\varepsilon} \le 1 + 12\sqrt{L_0\bar{R}^2/\varepsilon}$; at finisher levels $\sqrt{1 + 24L_0\bar{R}/(L_1\varepsilon)} \le 1 + 5\sqrt{L_0\bar{R}/(L_1\varepsilon)}$. Multiplying by the class counts (J^{sch}, S_Λ, K_{cert}) and using (25) ($\bar{\kappa} \le 5L_1\bar{R}$, hence $\bar{\kappa}/\sqrt{L_1\bar{R}} \le \sqrt{5}\,\sqrt{\bar{\kappa}}$),

$$\begin{aligned} N_{\text{det}} &\le 25\bar{K}\Big[J^{\text{sch}}\Big(2 + \sqrt{\tfrac{2L_0\bar{R}}{L_1\varepsilon}}\Big) + S_\Lambda\Big(1 + 12\sqrt{\tfrac{L_0\bar{R}^2}{\varepsilon}}\Big) \\ &\qquad + K_{\text{cert}}\Big(1 + 5\sqrt{\tfrac{L_0\bar{R}}{L_1\varepsilon}}\Big)\Big] \\ &= \widetilde{O}\Big(L_1R_0 + \bar{\kappa}^{1/2}\sqrt{\tfrac{L_0\bar{R}^2}{\varepsilon}}\Big). \end{aligned} \tag{36}$$

Statistical part. Since $\delta_j = \min\{\lambda_j s_j^2/2, \varepsilon/8\}$, we have $\frac{1}{\lambda_j\delta_j} \le \max\{\frac{2}{(\lambda_j s_j)^2}, \frac{8}{\lambda_j\varepsilon}\}$, and we bound $\lambda_j s_j$ per class: *Gradient-driven levels:* the two branches of s_j give $\lambda_j s_j \ge 4L_1\hat{G}_j \min\{\frac{R_0}{4J_{\text{tot}}}, \frac{\iota}{24L_0}\} \ge \frac{\hat{G}_j}{48J_{\text{tot}}}$ (using $L_1R_0 \ge \frac{1}{12}$ and $\iota/(24L_0) = \frac{1}{192L_1J_{\text{tot}}}$). Hence for *cold* gradient-driven levels ($\hat{G}_j \le \bar{G}$): $\frac{2}{(\lambda_j s_j)^2} \le \frac{2\cdot 96^2 J_{\text{tot}}^2\bar{R}^2}{\varepsilon^2}$ and $\frac{8}{\lambda_j\varepsilon} \le \frac{4\bar{R}}{L_1\varepsilon^2} \le \frac{16\bar{R}^2}{\varepsilon^2}$; for *hot* gradient-driven levels ($\hat{G}_j \ge \bar{G}$): $\frac{2}{(\lambda_j s_j)^2} \le \frac{4608L_1^2J_{\text{tot}}^2}{L_0^2}$ and $\frac{8}{\lambda_j\varepsilon} \le \frac{2}{L_0\varepsilon} \le \frac{4608L_1^2J_{\text{tot}}^2}{L_0^2}$ (the last step uses $2304(L_1\bar{R})^2J_{\text{tot}}^2 \ge 1$ when $\varepsilon \ge L_0\bar{R}^2$, and $\frac{2}{L_0\varepsilon} \le \frac{2\bar{R}^2}{\varepsilon^2}$ otherwise). *Floor levels:* $\lambda_j s_j \ge \frac{\varepsilon}{4\bar{R}^2}\min\{\frac{R_0}{8S_\Lambda}, \frac{1}{1152L_1S_\Lambda}\} \ge \frac{\varepsilon}{4608L_1\bar{R}^2S_\Lambda}$, so $\frac{2}{(\lambda_j s_j)^2} \le \frac{2\cdot 4608^2(L_1\bar{R})^2S_\Lambda^2\bar{R}^2}{\varepsilon^2}$ and $\frac{8}{\lambda_j\varepsilon} \le \frac{32\bar{R}^2}{\varepsilon^2}$. *Finisher levels:* $\lambda_{\text{cert}} s_j \ge \frac{L_1\varepsilon}{\bar{R}} \cdot \frac{1}{192L_1J_{\text{tot}}} = \frac{\varepsilon}{192J_{\text{tot}}\bar{R}}$, so

$$\tfrac{1}{\lambda\delta} \le \max\{\tfrac{2\cdot 192^2J_{\text{tot}}^2\bar{R}^2}{\varepsilon^2}, \tfrac{64L_1\bar{R}}{\lambda_{\text{cert}}\varepsilon} = \tfrac{64\bar{R}^2}{\varepsilon^2}\}. \tag{37}$$

Measurements: $\tau_j \ge \hat{G}_j/16 \ge \frac{\varepsilon}{32\bar{R}}$ always, and $\tau_j \ge \bar{G}/16$ at hot levels, so $B_j^{\text{est}} \le 1 + \frac{1024c_g^2\sigma^2\bar{R}^2\ell}{\varepsilon^2}$, improved to $1 + \frac{256c_g^2\sigma^2L_1^2\ell}{L_0^2}$ at hot levels. Summing over the classes ($\le J^{\text{sch}}$ cold and hot gradient-

driven levels, $\le S_\Lambda$ floor levels, $\le K_{\mathrm{cert}}$ finisher levels),

$$\begin{aligned} N_{\mathrm{stat}} \le 96 C_b \sigma^2 \ell \, \frac{\bar{R}^2}{\varepsilon^2} \Big[2{\cdot}10^4 J_{\mathrm{tot}}^3 + 5{\cdot}10^7 \bar{\kappa}^2 S_\Lambda^3 + 8{\cdot}10^4 J_{\mathrm{tot}}^2 K_{\mathrm{cert}} \Big] \\ + 5{\cdot}10^5 C_b \sigma^2 \ell \, \frac{L_1^2 J_{\mathrm{tot}}^3}{L_0^2} \\ + J_{\mathrm{tot}} \Big(1 + \frac{1024 c_g^2 \sigma^2 \bar{R}^2 \ell}{\varepsilon^2} \Big). \end{aligned} \tag{38}$$

Proof of Theorem 1. Correctness and termination. Proposition 4 shows that on $\mathcal{E}$ the run halts within J_{tot} levels and that its three exits give gaps at most $\varepsilon/2$, $5\varepsilon/8$, and $7\varepsilon/16$, respectively. The confidence budget of Appendix E.1 gives $\mathbb{P}[\mathcal{E}] \ge 1-\alpha$. The exact cost accounting established above is

$$N \le T_1 B_1 + N_{\mathrm{det}} + N_{\mathrm{stat}},$$

with the three terms given by (19), (36), and (38). Substitute the schedule constants: $J_{\mathrm{tot}} = \widetilde{O}(\bar{\kappa})$, $K_{\mathrm{cert}} = \widetilde{O}(\bar{\kappa})$, $S_G, S_\Lambda, \bar{K}, \ell = \widetilde{O}(1)$ (polylogarithmic). Then $N_{\mathrm{det}} = \widetilde{O}(L_1 R_0 + \bar{\kappa}^{1/2}\sqrt{L_0 \bar{R}^2/\varepsilon})$ by (36); the first bracket of (38) is $\widetilde{O}(\bar{\kappa}^3 \sigma^2 \bar{R}^2/\varepsilon^2)$, the second term is $\widetilde{O}(\bar{\kappa}^3 \sigma^2 L_1^2/L_0^2)$, the third is $\widetilde{O}(\bar{\kappa}\sigma^2\bar{R}^2/\varepsilon^2)$; and Proposition 2 gives $T_1 B_1 = \widetilde{O}\big(1 + L_1 R_0 + (1+L_1R_0)\sigma^2 L_1^2/L_0^2\big)$. Since $1 + L_1 R_0 \le \bar{\kappa}$ and $\bar{\kappa}^3\sigma^2 L_1^2/L_0^2$ absorbs the last Phase-I term, collecting the groups gives exactly the display of Theorem 1.

□

F Strongly convex restarts: proof of Theorem 2

The strongly convex wrapper and theorem are stated in the main paper; this section contains only their accounting and proof.

Default-certificate accounting for Remark 2. In default mode the round-k Phase I overhead has $q_k = L_{1,\mathrm{eff},k} R_k = \max\{q 2^{-k}, 1/12\}$, $q = L_1 R_0$. Hence $\sum_k q_k^2 = O(q^2 + K)$ and $\sum_k q_k^4 = O(q^4 + K)$. The floor-active contribution $(1+q_k)^4 \sigma^2 L_{1,\mathrm{eff},k}^2 / L_0^2$ is absorbed into the round's statistical term by Lemma 16; the remaining first-round contributions are exactly the two ε-independent terms stated in Remark 2.

Lemma 16 (Effective-parameter absorption)**.** *Let $R_k = R_0 2^{-k}$, $\bar{R}_k := 3R_k$, $L_{1,\mathrm{eff},k} := \max\{L_1, 1/(12R_k)\}$, $\bar{\kappa}_{\mathrm{eff},k} := 1 + 3L_{1,\mathrm{eff},k} R_k$, and let round k run in the nontrivial regime $\varepsilon_k < \Delta_{\mathrm{cert},k} \le \Delta_{\mathrm{def},k} := \frac{L_0}{L_{1,\mathrm{eff},k}^2} \psi\big(L_{1,\mathrm{eff},k} R_k\big)$. Then on every floor-active round $k \in \mathcal{K}_{\mathrm{fl}} := \{k : L_{1,\mathrm{eff},k} = 1/(12R_k)\}$:*

$$\begin{aligned} \Delta_{\mathrm{def},k} = 144 L_0 R_k^2 \, \psi\big(\tfrac{1}{12}\big) < 0.52 \, L_0 R_k^2, \qquad \textit{hence} \\ \varepsilon_k^2 < 0.27 \, L_0^2 R_k^4, \\ \frac{\sigma^2}{L_0^2 R_k^2} < 0.03 \, \frac{\sigma^2 \bar{R}_k^2}{\varepsilon_k^2}. \end{aligned} \tag{39}$$

consequently each of $\sigma^2 L_{1,\mathrm{eff},k}^3 R_k / L_0^2$, $\bar{\kappa}_{\mathrm{eff},k}^3 \sigma^2 L_{1,\mathrm{eff},k}^2 / L_0^2$, and $\sigma^2 L_{1,\mathrm{eff},k}^2 J_{\mathrm{tot},k}^3 / L_0^2$ is at most an absolute-constant (resp. polylogarithmic) multiple of the round's own statistical term $\bar{\kappa}_{\mathrm{eff},k}^3 \sigma^2 \bar{R}_k^2/\varepsilon_k^2$.

Proof. The hypothesis $\Delta_{\mathrm{cert},k} \le \Delta_{\mathrm{def},k}$ holds on every round by the $\min\{\varepsilon_{k-1}, \Delta_{\mathrm{def},k}\}$ construction of the restart wrapper in the main paper (in user mode, $\min\{\Delta_{\mathrm{user}}, \varepsilon_{k-1}, \Delta_{\mathrm{def},k}\}$), and the regime is genuine: on a nontrivial round running on the pass-through certificate one has $\varepsilon_k < \varepsilon_{k-1}$, which forces $\varepsilon_{k-1} = \frac{\mu R_{k-1}^2}{8} = \frac{\mu R_k^2}{2} \le \frac{L_0}{2} R_k^2 < 0.52 \, L_0 R_k^2$, since $\mu \le L_0$ (Remark 3). On $k \in \mathcal{K}_{\mathrm{fl}}$ one has $L_{1,\mathrm{eff},k} R_k = \frac{1}{12}$, so $\Delta_{\mathrm{def},k} = \frac{L_0}{L_{1,\mathrm{eff},k}^2} \psi\big(\frac{1}{12}\big) = 144 L_0 R_k^2 \psi\big(\frac{1}{12}\big)$, and

$\psi\big(\frac{1}{12}\big) = e^{1/12} - 1 - \frac{1}{12} \approx 0.0035707$ gives $144\,\psi\big(\frac{1}{12}\big) \approx 0.51418 < 0.52$. The nontrivial regime then yields $\varepsilon_k < \Delta_{\mathrm{def},k}$, hence $\varepsilon_k^2 < 0.5142^2\, L_0^2 R_k^4 \le 0.27\, L_0^2 R_k^4$, and

$$\frac{\sigma^2}{L_0^2 R_k^2} < 0.2644\,\frac{\sigma^2 R_k^2}{\varepsilon_k^2} = \frac{0.2644}{9}\,\frac{\sigma^2 \bar{R}_k^2}{\varepsilon_k^2} < 0.03\,\frac{\sigma^2 \bar{R}_k^2}{\varepsilon_k^2},$$

using $\bar{R}_k = 3R_k$. On floor-active rounds $\bar{\kappa}_{\mathrm{eff},k} = 1 + \frac{1}{4} = \frac{5}{4}$ and $L_{1,\mathrm{eff},k}^3 R_k = \frac{1}{1728 R_k^2}$ (since $1728 = 12^3$), so $\sigma^2 L_{1,\mathrm{eff},k}^3 R_k / L_0^2 = \sigma^2/(1728 L_0^2 R_k^2)$ and $\bar{\kappa}_{\mathrm{eff},k}^3 \sigma^2 L_{1,\mathrm{eff},k}^2 / L_0^2 = \big(\frac{5}{4}\big)^3 \sigma^2/(144 L_0^2 R_k^2)$ are absolute-constant multiples of $\sigma^2/(L_0^2 R_k^2)$, hence of the round's statistical term by (39); and $J_{\mathrm{tot},k} \le 114 \bar{\kappa}_{\mathrm{eff},k} S_G (S_\Lambda + 4)$ is polylogarithmic there, so the hot-level term is a polylogarithmic multiple of the same. □

Proof of Theorem 2. Work on the intersection of the K good events (probability $\ge 1 - \alpha$). *Correctness.* By induction, $\|x_k - x^\star\| \le R_k$: the base is Assumption 1. For every $k \le K - 2$ one has $\frac{\mu R_k^2}{8} > \varepsilon$ (since $k \le \big\lceil \log_4 \frac{\mu R_0^2}{8\varepsilon} \big\rceil - 1 < \log_4 \frac{\mu R_0^2}{8\varepsilon}$ by the choice of K), so round k targets $\varepsilon_k = \frac{\mu R_k^2}{8}$ and, given $\|x_k - x^\star\| \le R_k$, returns $f(x_{k+1}) - f^\star \le \frac{\mu R_k^2}{8}$; by μ-strong convexity $\|x_{k+1} - x^\star\|^2 \le \frac{2\varepsilon_k}{\mu} = \frac{R_k^2}{4} = R_{k+1}^2$, closing the induction up to $k = K - 1$ and validating the last round's inputs. By the choice of K, $\frac{\mu R_{K-1}^2}{8} \le \varepsilon$, so the last round targets $\varepsilon_{K-1} = \varepsilon$ and the output satisfies $f(x_K) - f^\star \le \varepsilon$.

Cost. Apply Theorem 1 per round with $(R_k, \varepsilon_k, \alpha/K)$, $\bar{R}_k = 3R_k$, $\bar{\kappa}_k = 1 + 3L_1 R_k$, and sum. Round k re-derives its own effective parameter $L_{1,\mathrm{eff},k} := \max\{L_1, 1/(4\bar{R}_k)\} = \max\{L_1, 1/(12R_k)\}$ (Algorithm 1); write $\bar{\kappa}_{\mathrm{eff},k} := 1 + 3L_{1,\mathrm{eff},k} R_k$ and split the rounds into $\mathcal{K}_0 := \{k : L_{1,\mathrm{eff},k} = L_1\}$ and $\mathcal{K}_{\mathrm{fl}} := \{k : L_{1,\mathrm{eff},k} = 1/(12R_k)\}$. Since $L_{1,\mathrm{eff},k} R_k \le L_1 R_k + \frac{1}{12}$ and $\bar{\kappa}_{\mathrm{eff},k} \le \frac{5}{4}\bar{\kappa}_k$, every term depending on $L_{1,\mathrm{eff},k}$ only through $\bar{\kappa}_{\mathrm{eff},k}$ or through the linear product $L_{1,\mathrm{eff},k} R_k$ changes by at most the constant $\big(\frac{5}{4}\big)^3$ or the additive $\sum_k L_{1,\mathrm{eff},k} R_k \le 2L_1 R_0 + \frac{K}{12}$. The remaining families—the Phase I burn-in $\sigma^2 L_{1,\mathrm{eff},k}^3 R_k / L_0^2$, the burn-in $\bar{\kappa}_{\mathrm{eff},k}^3 \sigma^2 L_{1,\mathrm{eff},k}^2 / L_0^2$, and the hot-level term $\sigma^2 L_{1,\mathrm{eff},k}^2 J_{\mathrm{tot},k}^3 / L_0^2$ of (38)—coincide on $\mathcal{K}_0$ with the L_1-versions summed below, while on $\mathcal{K}_{\mathrm{fl}}$ each is $O\big(\sigma^2/(L_0^2 R_k^2)\big)$ up to polylogarithmic factors, since $\bar{\kappa}_{\mathrm{eff},k} = \frac{5}{4}$ and $L_{1,\mathrm{eff},k}^3 R_k = 1/(1728 R_k^2)$ there; by Lemma 16 each such term is absorbed into the statistical term of its own round (a floor-active round in the trivial regime $\varepsilon_k \ge \Delta_{\mathrm{cert},k}$ makes no oracle calls at all, so there is nothing to absorb). We therefore carry out the sums with L_1. *Deterministic part:* $\sqrt{L_0 \bar{R}_k^2/\varepsilon_k} \le \sqrt{72 L_0/\mu}$ for every k (with equality on the $\mu R_k^2/8$-rounds; on the last round $\bar{R}_{K-1}^2 \le 9 \cdot \frac{8\varepsilon}{\mu}$ and $\varepsilon_{K-1} = \varepsilon$), so

$$\sum_k \bar{\kappa}_k^{1/2} \sqrt{\tfrac{L_0 \bar{R}_k^2}{\varepsilon_k}} \le \sqrt{\tfrac{72 L_0}{\mu}} \Big(K + \sqrt{3L_1 R_0} \sum_k 2^{-k/2} \Big) \le \sqrt{\tfrac{72 L_0}{\mu}} \big(K + 4\sqrt{3L_1 R_0} \big), \tag{40}$$

and $\sqrt{L_1 R_0} \le 1 + L_1 R_0$, while $\sum_k L_1 R_k \le 2L_1 R_0$; using $K = \widetilde{O}\big(\log_+(\mu R_0^2/\varepsilon)\big)$, this gives the group $\widetilde{O}\big((1 + L_1 R_0)(1 + \sqrt{L_0/\mu}) + \sqrt{L_0/\mu}\,\log_+(\mu R_0^2/\varepsilon)\big)$. *Statistical part:* with $a_k := 3L_1 R_k$, the $\mu R_k^2/8$-rounds $k \le K - 2$ (present only when $K \ge 2$) have $\bar{R}_k^2/\varepsilon_k^2 = 576/(\mu^2 R_k^2)$, so

$$\begin{aligned}
\sum_{k \le K-2} \bar{\kappa}_k^3 \frac{\sigma^2 \bar{R}_k^2}{\varepsilon_k^2} &= \frac{576\sigma^2}{\mu^2} \sum_{k \le K-2} \frac{(1 + a_k)^3}{R_k^2} \\
&\le \frac{576\sigma^2}{\mu^2} \Bigg[\underbrace{\sum_{k \le K-2} \tfrac{1}{R_k^2}}_{\le \frac{4}{3}/R_{K-2}^2 \le \mu/(6\varepsilon)} + \underbrace{9L_1 \sum_{k \le K-2} \tfrac{1}{R_k}}_{\le 13 L_1 \sqrt{\mu/\varepsilon}} \\
&\qquad + 27 L_1^2 K + \underbrace{27 L_1^3 \sum_{k \le K-2} R_k}_{\le 54 L_1^3 R_0} \Bigg],
\end{aligned}$$

where the geometric sums are dominated by their last term and $R_{K-2}^2 > \frac{8\varepsilon}{\mu}$ (since $K-2 < \log_4 \frac{\mu R_0^2}{8\varepsilon}$); the last round costs

$$\bar{\kappa}_{K-1}^3 \frac{\sigma^2 \bar{R}_{K-1}^2}{\varepsilon^2} \le \frac{9 R_{K-1}^2 \bar{\kappa}_{K-1}^3 \sigma^2}{\varepsilon^2} \le \frac{72\, \bar{\kappa}_{K-1}^3 \sigma^2}{\mu\varepsilon}$$

since $R_{K-1}^2 \le \frac{8\varepsilon}{\mu}$. The total is $\widetilde{O}\big((1 + L_1\sqrt{\varepsilon/\mu})^3 \frac{\sigma^2}{\mu\varepsilon} + \frac{\sigma^2 L_1^3 R_0}{\mu^2}\big)$ after matching the summands with the expansion of $(1 + L_1\sqrt{\varepsilon/\mu})^3/(\mu\varepsilon)$ (the $27L_1^2 K$ summand contributes $\widetilde{O}(\sigma^2 L_1^2/\mu^2)$, absorbed with a logarithmic factor; the L_1^3 summand is the ε-independent restart burn-in $\sigma^2 L_1^3 R_0/\mu^2$ stated in the theorem; and $\bar{\kappa}_{K-1} \le 1 + 3L_1\sqrt{8\varepsilon/\mu}$, so the last-round term enters the first group). *Burn-ins:* $\sum_k \sigma^2 L_1^3 R_k/L_0^2 \le 2\sigma^2 L_1^3 R_0/L_0^2 \le 2\sigma^2 L_1^3 R_0/\mu^2$ (using $\mu \le L_0$, Remark 3) and, using $(1+a)^3 \le 4(1+a^3)$,

$$\sum_k \bar{\kappa}_k^3 \tfrac{\sigma^2 L_1^2}{L_0^2} \le \big(4K + 124(L_1 R_0)^3\big) \tfrac{\sigma^2 L_1^2}{L_0^2} = \widetilde{O}\big(\bar{\kappa}^3 \tfrac{\sigma^2 L_1^2}{L_0^2}\big). \tag{41}$$

Collecting the groups yields the display of Theorem 2. Finally, when $\varepsilon \le \frac{\mu}{36 L_1^2}$ we have $L_1\sqrt{\varepsilon/\mu} \le \frac{1}{6}$, so the leading statistical term is at most $1.6 \cdot \frac{\sigma^2}{\mu\varepsilon}$ up to the logarithmic factor ℓ, matching the classical $\Omega(\sigma^2/(\mu\varepsilon))$ information-theoretic lower bound. □

G Supplementary remarks and comparisons

The following supplemental remarks record specializations and comparisons that are not needed to state or run the method.

Remark 6 (Details of the smooth-case recovery)**.** *The smooth-case rates are stated after Theorems 1–2 in the main paper. Here are the details behind that specialization. For $L_1 = 0$ the substitution convention of Appendix A applies ($L_{1,\mathrm{eff}} = 1/(4\bar{R})$, so $\bar{\kappa} \le \frac{5}{4}$) and* Phase I *terminates immediately. In either certificate regime, under the substitution $L_{1,\mathrm{eff}} R_0 = \frac{1}{12}$, the nontrivial regime $\varepsilon < \Delta_{\mathrm{cert}}$ gives $\Delta_{\mathrm{cert}} \le \Delta_{\mathrm{def}}(L_{1,\mathrm{eff}}) = 16 L_0 \bar{R}^2 \psi(\frac{1}{12}) \approx 0.057 L_0 \bar{R}^2$—note that $\Delta_{\mathrm{def}} = O(L_0 R_0^2)$ holds only under the $L_{1,\mathrm{eff}}$-substitution, or trivially at $L_1 = 0$—hence both effective-L_1 burn-ins are absorbed by the statistical term $\sigma^2\bar{R}^2/\varepsilon^2$ with absolute constants ≤ 0.03: from $\varepsilon^2 < 0.0572^2 L_0^2 \bar{R}^4 \le 0.003264\, L_0^2 \bar{R}^4$ one gets $\sigma^2/(L_0^2 R_0^2) = 9\sigma^2/(L_0^2\bar{R}^2) \le 9 \cdot 0.0033\, \sigma^2\bar{R}^2/\varepsilon^2 \le 0.03\, \sigma^2\bar{R}^2/\varepsilon^2$ and $\bar{\kappa}_{\mathrm{eff}}^3 \sigma^2 L_{1,\mathrm{eff}}^2/L_0^2 = \frac{125}{1024}\sigma^2/(L_0^2\bar{R}^2) \le 4 \cdot 10^{-4}\, \sigma^2\bar{R}^2/\varepsilon^2$ with $\bar{\kappa}_{\mathrm{eff}} := 1 + L_{1,\mathrm{eff}}\bar{R} = \frac{5}{4}$; likewise the* Phase I *logarithm is $O(1)$ since $\log_+\big(2L_{1,\mathrm{eff}}^2 \Delta_{\mathrm{cert}}/L_0\big) \le \log_+\big(2\psi(\frac{1}{12})\big) = 1$. The statistical terms are unimprovable in their ε-dependence for any $L_1 \ge 0$ by the classical lower bounds (Nemirovsky & Yudin, 1983), since (L_0, L_1)-smooth functions include L_0-smooth ones; we are not aware of lower bounds quantifying the necessary dependence on $\bar{\kappa} = 1 + L_1\bar{R}$ in the stochastic setting, and our $\bar{\kappa}^{1/2}$- and $\bar{\kappa}^3$-factors should be read as upper-bound artifacts whose necessity is open (Section G.1).*

Remark 7 (Deterministic oracle)**.** *For $\sigma = 0$ the bound of Theorem 1 becomes $\widetilde{O}(1 + L_1 R_0 + \bar{\kappa}^{1/2}\sqrt{L_0\bar{R}^2/\varepsilon})$ (with Δ_{cert} as an external certificate; under the default certificate the $1 + L_1 R_0$ term is $\Theta((1+q)^2)$, Corollary 2). This has the same "accelerated term + $L_1 R$-burn-in" structure as the deterministic accelerated methods of Vankov et al. (2025); Gorbunov et al. (2025), but is* not *state of the art in that setting: it carries an extra $\bar{\kappa}^{1/2}$-factor on the accelerated term relative to the $(\nu+1)\big[\sqrt{L_0 R^2/\varepsilon} + (L_1 R)^{2/3}\log(2F_0/\varepsilon)\big]$ bound of Vankov et al. (2025) (where ν is the per-iteration line-search cost) and to the $\sqrt{L_0(1 + L_1 R e^{L_1 R})R^2/\varepsilon}$ bound of Gorbunov et al. (2025), whose exponential factor affects only the regime $\varepsilon \gtrsim L_0 R^2$. Our contribution is the stochastic setting; specializing it to $\sigma = 0$ is not the intended use.*

Remark 8 (Criterion $\|\nabla f\| \le \varepsilon_{\mathrm{g}}$)**.** *Run with the gradient-based stopping rule only,* Phase II *outputs $\hat{x}$ with $\|\nabla f(\hat{x})\| \le \varepsilon_{\mathrm{g}}$ using $\widetilde{O}\big(1 + L_1 R_0 + L_1\bar{R}\sqrt{1 + L_0/(L_1\varepsilon_{\mathrm{g}})} + \bar{\kappa}^3\sigma^2/\varepsilon_{\mathrm{g}}^2 \cdot L_1^2\bar{R}^2\big)$ calls (with Δ_{cert} as an external certificate; under the default certificate the $1 + L_1 R_0$ entry term becomes $\Theta((1+q)^2)$, Corollary 2); for $L_1 = 0$ (via the $L_{1,\mathrm{eff}}$ substitution) and moderate accuracies this is within logarithmic factors of the (smooth-case optimal) gradient-norm complexity of Foster et al. (2019), whose recursive-regularization architecture our levels generalize.*

Remark 9 (Scope relative to stochastic acceleration and ball methods)**.** *The convex accuracy exponents of Theorem 1 are not claimed as the first such exponents under generalized smoothness. The RSAG analysis of Yu et al. (2025) has a $\widetilde{O}(T^{-1/2})$ general-noise gap rate and an accelerated dependence in its low-noise regime; our additive oracle is likewise a specialization of that work's affine-variance family, not an incomparable model.*

Ball localization is also an established ingredient: see the ball optimization oracle and its acceleration in Carmon et al. (2020), the stochastic ReSQue construction of Carmon et al. (2023), and the Broximal method of Gruntkowska et al. (2025). The concurrent 2026 TRPPM preprint—subsequent to those earlier ball frameworks—by Li et al. (2026) studies the same formal ball-constrained proximal objective, but in a deterministic trust-region regime that seeks an active constraint. Here the regularization floor instead keeps the exact prox in the interior, and the ball contains every query of the approximate stochastic solver. The scoped contribution of ARC-SG is this query-by-query certified geometry, explicit (L_0, L_1) parameter accounting, and the separate strongly convex restart guarantee, rather than the ball-proximal form or the convex accuracy exponents in isolation.

G.1 On the $\bar{\kappa}$-factors, and what is known to be optimal

Where the factors come from. The factors $\bar{\kappa}^{1/2}$ and $\bar{\kappa}^3$ in Theorem 1 have a single, well-identified origin. All queries of level j live in a ball of radius $O(1/L_1)$; consequently, information must "hop" between balls, and the momentum of the inner accelerated method does not carry across hops. The $\bar{\kappa}^{1/2}$-factor on the optimization term is the price of restarting acceleration $\widetilde{O}(L_1\bar{R})$ times at the finest floor scale; the $\bar{\kappa}^3$-factor on the statistical term is the price of maintaining, across $\widetilde{O}(L_1\bar{R})$ noisy levels, the additive precision budget $\sum_j s_j \lesssim 1/L_1$ that protects the gradient-cap invariant against noise-induced drift (Lemma 13), at a statistical cost $\propto s_j^{-2}$ per level. It is worth being explicit that the second mechanism makes the gradient *measurements* the dominant cost in the worst case: the per-level anchors are individually cheap relative to a level's optimization work, but their total is what carries the $\bar{\kappa}^3$.

Dimensional analysis. Write $q := L_1\bar{R}$ (dimensionless), $A := L_0\bar{R}^2/\varepsilon$ and $V := \sigma^2\bar{R}^2/\varepsilon^2$. Theorem 1 then reads $\widetilde{O}(1 + q + \bar{\kappa}^{1/2}\sqrt{A} + \bar{\kappa}^3 V)$ with $\bar{\kappa} = 1 + q$—the additive entry price $1 + q$ under an externally supplied gap certificate; under the default gap certificate the proved entry price is $\Theta((1+q)^2)$ (Corollary 2)—whereas the target of Open Problem 1 is $\widetilde{O}(\Phi(q) + \sqrt{A} + V)$ for some additive entry price Φ. The distinction between an *additive* $\Phi(q)$ and a *multiplicative* $\mathrm{poly}(q)$ in front of $\sqrt{A}$ and V is the substantive one, and it is achievable in neighbouring settings: the deterministic method of Vankov et al. (2025) has the additive structure $\sqrt{A} + q^{2/3}\log(F_0/\varepsilon)$, Tyurin (2026) obtain, at small ε, a leading accelerated term free of nonconstant multiplicative q-dependence, and in the non-accelerated stochastic regime Gaash et al. (2025) already achieve smooth-optimal leading terms $L_0R^2/\varepsilon + \sigma^2R^2/\varepsilon^2$ with L_1 entering only additively. Separation is therefore possible in principle, and our bound does not achieve it.

What is and is not proved optimal. To delineate precisely. *Optimal:* the ε-exponents of both the optimization and the statistical terms, since (L_0, L_1)-smooth functions contain the L_0-smooth ones and the classical lower bounds (Nemirovsky & Yudin, 1983) apply on that subclass. *Not proved optimal, and probably not optimal:* the multiplicative factors $\bar{\kappa}^{1/2}$ and $\bar{\kappa}^3$; the form L_1R_0 of the burn-in; the threshold $\varepsilon \le \mu/(36L_1^2)$ in Theorem 2; and, in the overparameterized regime, the dependence on ρ and the cube $(1 + L_0/\mu)^3$. We are not aware of *any* stochastic lower bound that certifies growth in q of the leading constants, so all q-dependence beyond the smooth-case bounds should be read as an upper-bound artifact. We also correct a comparison made in the previous version of this paper: our burn-in L_1R_0 does *not* match the deterministic burn-in of Vankov et al. (2025) "up to logarithms"—theirs is $(L_1R)^{2/3}\log(2F_0/\varepsilon)$, a genuinely smaller polynomial—and after Tyurin (2026) the deterministic state of the art is further away still. Our deterministic specialization ($\sigma = 0$) is therefore strictly weaker than the best deterministic methods; the contribution is the stochastic setting.

A caveat on parameterization. Lower-bound statements in this model must also fix *which* pair (L_0, L_1) is meant, because the pair is not unique: if f is (L_0, L_1)-smooth it is also (L_0', L_1')-smooth

Result	Convexity	Smooth.	Oracle	Known params
Thm. 1	convex	A2	A3	L_0, L_1, σ, R_0
Thm. 2	μ-str. cvx	A2	A3	$+\,\mu$
Thm. 3†	μ-str. cvx	A2 ind.	A4 a.s.	L_0, L_1, ρ, R_0, μ
Thm. 4†	convex	A2 ind.	A4 a.s.	L_0, L_1, ρ, R_0

Table 2: Assumptions per theorem. "A2 ind." means every f_ξ satisfies Assumption 2 individually; "A4 a.s." is the almost-sure strong-growth variant (43) (a.s. = almost surely, used throughout). Assumption 1 and a valid upper bound Δ_{cert} (default: Lemma 7) are required throughout; all guarantees are high-probability. Rows marked † concern the overparameterized regime and are stated and proved in Appendix H.

for every $L_0' \geq L_0$, $L_1' \geq L_1$, and the admissible set typically contains a nontrivial Pareto frontier. A meaningful lower bound with explicit $q = L_1 R$ dependence should either be stated for a canonical (e.g. Pareto-minimal) pair, or with the complexity defined as the infimum of the bound over all admissible pairs. Our upper bounds hold for every admissible pair and hence for the best one; but a matching lower bound requires committing to a canonical parameterization, and we regard the formulation of that class as part of the open problem.

Remark 10 (Scaling invariance of (q, A, V))**.** *The groups q, A, V are the natural dimensionless coordinates of the problem: they are invariant under the two-parameter rescaling that acts on (1) by a domain dilation $x \mapsto \alpha x$ and a range scaling $f \mapsto \beta f$ ($\alpha, \beta > 0$). Indeed, writing $g(x') := \beta f(\alpha x')$ gives $\nabla g(x') = \alpha\beta\, \nabla f(\alpha x')$ and $\nabla^2 g(x') = \alpha^2 \beta\, \nabla^2 f(\alpha x')$, so g is $(\alpha^2\beta L_0,\ \alpha L_1)$-generalized smooth, with initial distance $\bar{R}/\alpha$, target gap $\beta\varepsilon$, and—taking the g-oracle to be $\alpha\beta$ times the f-oracle—sub-Gaussian scale $\alpha\beta\sigma$; convexity and unbiasedness are preserved. Substituting these into $q = L_1\bar{R}$, $A = L_0\bar{R}^2/\varepsilon$, $V = \sigma^2\bar{R}^2/\varepsilon^2$ leaves all three unchanged: $q^g = q$, $A^g = A$, $V^g = V$. Consequently both our upper bounds and any matching lower bound can—and, for a canonical statement, should—be written purely in (q, A, V); this is the form in which Open Problem 1 is posed.*

Open Problem 1. *Under Assumptions 1–3, with $q = L_1\bar{R}$, $A = L_0\bar{R}^2/\varepsilon$, $V = \sigma^2\bar{R}^2/\varepsilon^2$: (i) is there a high-probability accelerated stochastic method with complexity $\widetilde{O}(\Phi(q) + \sqrt{A} + V)$, i.e. with* no *multiplicative q-factor on the leading terms? (ii) what is the minimal additive entry price $\Phi(q)$—is it q, $q^{2/3}$, $\log q$, or something else? (iii) can one prove a stochastic lower bound with explicit dependence on $q > 0$ for a canonically parameterized class of (L_0, L_1)-smooth functions? A positive answer to (i) would likely require carrying stochastic momentum across smoothness cells, e.g. via a high-probability analysis of clipped SSTM (Gorbunov et al., 2020; Sadiev et al., 2023) with (L_0, L_1)-adaptive clipping levels, or a redistribution of the drift budget that avoids spending precision uniformly across levels.*

Table 2 collects the assumptions and oracle models required by each theorem.

H Strong-growth and interpolation extensions

This section gives the complete treatment of the overparameterized regime: the setting and its relation to exact interpolation (Appendix H.1), the main guarantees (Appendix H.2), the hop-and-solve algorithm and its analysis (Appendices H.3–H.4), the obstruction behind Open Problem 2 (Appendix H.5), and the corresponding numerical illustration (Appendix J.1). It uses the notation and baseline assumptions of the main paper. This appendix assumes $L_1 > 0$ (equivalently, it applies the $L_{1,\mathrm{eff}}$ convention of Algorithm 1).

H.1 Setting: strong growth under generalized smoothness

Modern overparameterized models often *interpolate*: a single x^{int} minimizes every f_ξ simultaneously. In the smooth case this makes SGD, and with momentum also accelerated methods, converge at deterministic-like rates (Schmidt & Le Roux, 2013; Ma et al., 2018; Vaswani et al., 2019; Liu & Belkin, 2020). The generalized-smooth analogue that our analysis actually uses is the following strong growth condition.

Method / reference	Setting	Smoothness	Noise	Oracle complexity (leading terms)
(A) Deterministic oracle (exact gradients)				
AGMsDR, line search (Vankov et al., 2025)	convex	(L_0, L_1)	—	$(\nu{+}1)[\sqrt{\frac{L_0R^2}{\varepsilon}} + (L_1R)^{2/3}\log\frac{2F_0}{\varepsilon}]$
Clipped AGD (Gorbunov et al., 2025)	convex	(L_0, L_1)	—	$\sqrt{\frac{L_0(1+L_1R\,e^{L_1R})R^2}{\varepsilon}}$
Near-optimal AGM (Tyurin, 2026)	convex	(L_0, L_1)	—	$\sqrt{\frac{L_0R^2}{\varepsilon}}$ + additive L_1-terms
ARC-SG specialization ($\sigma = 0$)	convex	(L_0, L_1)	—	$1 + L_1R_0 + \bar{\kappa}^{1/2}\sqrt{\frac{L_0R^2}{\varepsilon}}$
(B) Stochastic first-order oracle				
SGD (Nemirovski et al., 2009)	convex, in exp.	L	σ^2	$\frac{LR^2}{\varepsilon} + \frac{\sigma^2R^2}{\varepsilon^2}$
AC-SA (Lan, 2012; Ghadimi & Lan, 2012)	convex, in exp.	L	σ^2	$\sqrt{\frac{LR^2}{\varepsilon}} + \frac{\sigma^2R^2}{\varepsilon^2}$
ClipSSTM (Gorbunov et al., 2020)	convex, high prob.	L	heavy tails	$\sqrt{\frac{LR^2}{\varepsilon}} + \frac{\sigma^2R^2}{\varepsilon^2}$
Recursive reg. (Foster et al., 2019)	convex, $\|\nabla f\| \le \varepsilon_g$	L	σ^2	$\sqrt{\frac{LF_0}{\varepsilon_g^2}} + \frac{\sigma^2}{\varepsilon_g^2}$
SGD, (L_0, L_1)-steps (Lobanov & Gasnikov, 2025)	convex, in exp.	(L_0, L_1)	σ^2	$\frac{L_0\bar{R}^2}{\varepsilon} + \frac{\sigma^2\bar{R}^2}{\varepsilon^2}$ + burn-in(L_1)
Clipped SGD (Gaash et al., 2025)	convex, high prob.	(L_0, L_1)	norm-sub-Gaussian	$\frac{L_0R^2}{\varepsilon} + \frac{\sigma^2\bar{R}^2}{\varepsilon^2}$ + additive terms
Acc. (RSAG) (Yu et al., 2025)	convex, high prob.	generalized	affine var.	statistical ε^{-2} (gen. noise); accel. opt. $\varepsilon^{-1/2}$ (low noise)‡
ARC-SG (Thm. 1)	convex, high prob.	(L_0, L_1)	norm-sub-Gaussian	$1 + L_1R_0 + \bar{\kappa}^{\frac{1}{2}}\sqrt{\frac{L_0R^2}{\varepsilon}} + \bar{\kappa}^3\frac{\sigma^2\bar{R}^2}{\varepsilon^2}$
ARC-SG (Thm. 2)	μ-str. cvx, high prob.	(L_0, L_1)	norm-sub-Gaussian	$(1 + L_1R_0)\left(1 + \sqrt{\frac{L_0}{\mu}}\right) + \sqrt{\frac{L_0}{\mu}}\log_+\left(\frac{\mu R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$ $(\varepsilon \le \frac{\mu}{36L_1^2})$

Table 3: Complexity comparison, up to absolute constants and polylogarithmic factors ($\widetilde{O}$-notation); R denotes the respective initial-distance quantity of each work, $F_0 = f(x^0) - f^\star$, $\bar{\kappa} = 1 + L_1\bar{R}$, $\bar{R} = 3R_0$. The table is split by *oracle model*: panel (A) assumes exact gradients, panel (B) a stochastic first-order oracle; entries *across* panels are not comparable, and within panel (B) the guarantee type (in expectation versus high probability) and parameter knowledge differ line by line. In the AGMsDR row, ν bounds the number of additional oracle calls of the one-dimensional line search per iteration; arithmetic costs of line searches and projections are not counted anywhere in the table. The additive terms of Gaash et al. (2025), the burn-in of Lobanov & Gasnikov (2025), and the additive L_1-terms of Tyurin (2026) depend on their respective problem parameters and initial gaps, and we refer to the cited works for their exact form. ARC-SG differs from the other stochastic rows in its certified-query geometry, explicit (L_0, L_1) accounting, and strongly convex restart theorem. For $L_1 = 0$ our bounds recover the smooth-case complexities of the first two rows of panel (B), whose ε-exponents are optimal. Our deterministic specialization is included only for orientation: it is *not* state of the art in panel (A) (see the remarks after Theorem 2). The ARC-SG rows are printed without the ε-independent burn-in $\widetilde{O}((1 + L_1R_0)\sigma^2L_1^2/L_0^2 + \bar{\kappa}^3\sigma^2L_1^2/L_0^2)$ of Theorem 1 (for Theorem 2, $\widetilde{O}(\sigma^2L_1^3R_0/\mu^2 + \bar{\kappa}^3\sigma^2L_1^2/L_0^2)$), which is not dominated by the statistical term over the full accuracy range; and their $\widetilde{O}$-forms treat the initial-gap bound Δ_{cert} as an externally supplied certificate—under the default gap certificate of Lemma 7, with $q := L_1R_0$, the Phase I entry terms scale as $(1 + q)^2$ in the deterministic part and as $(1 + q)^4\sigma^2L_1^2/L_0^2$ in the statistical burn-in (Corollary 2 and Remarks 1–2). Method abbreviations: AGMsDR is the Accelerated Gradient Method with Small-Dimensional Relaxation of Vankov et al. (2025); AGD is accelerated gradient descent; AGM is accelerated gradient method; AC-SA is the Accelerated Stochastic Approximation; ClipSSTM is the clipped stochastic similar-triangles method (SSTM) of Gorbunov et al. (2020). Cell shorthands: "in exp." = in expectation; "high prob." = high probability; "μ-str. cvx" = μ-strongly convex; "sub-Gauss." = sub-Gaussian; "gen. noise" = general noise; "accel. opt." = accelerated optimization; "affine var." = affine variance; "Recursive reg." = recursive regularization.

Assumption 4 (SGC-GS (strong growth condition under generalized smoothness))**.** *Each f_ξ is convex, differentiable and (L_0, L_1)-smooth in the sense of (2); the oracle returns $g(x, \xi) = \nabla f_\xi(x)$; and there exists $\rho \ge 1$ such that for all $x \in \mathbb{R}^d$,*

$$\mathbb{E}_\xi \left\|\nabla f_\xi(x)\right\|^2 \;\le\; \rho \left\|\nabla f(x)\right\|^2. \tag{42}$$

For the theorems below we use the almost-sure variant

$$\left\|\nabla f_\xi(x)\right\| \le \sqrt{\rho}\left\|\nabla f(x)\right\| \qquad \text{for a.e. } \xi \text{ and all } x \in \mathbb{R}^d, \tag{43}$$

which implies (42); Remark 13 discusses the obstructions to working under (42) alone.

Exact interpolation is a *motivating special case* of Assumption 4, not an equivalent condition; the next lemma makes the logical relation precise. To avoid the ambiguity of writing "$f_\xi(x^{\mathrm{int}}) = \min f_\xi$", we state the hypothesis explicitly: the minimum is over the decision variable $x \in \mathbb{R}^d$, for a fixed realization ξ, and the minimizing point is required to be the *same* x^{int} for almost every ξ.

Definition 1 (Exact interpolation). *The family* $\{f_\xi\}$ interpolates *at* $x^{\mathrm{int}} \in \mathbb{R}^d$ *if*

$$f_\xi(x^{\mathrm{int}}) = \inf_{x\in\mathbb{R}^d} f_\xi(x) \qquad \text{for a.e. } \xi, \tag{44}$$

i.e., the common point x^{int} *minimizes every individual loss* f_ξ *over the decision space. Equivalently, since each* f_ξ *is convex and differentiable,* $\nabla f_\xi(x^{\mathrm{int}}) = 0$ *for a.e.* ξ. *Note that* (44) *forces* $x^{\mathrm{int}} \in X^\star$ *and* $f^\star = \mathbb{E}_\xi f_\xi(x^{\mathrm{int}})$.

Lemma 17 (Interpolation implies generalized weak growth). *Let each* f_ξ *be convex, differentiable and* (L_0, L_1)*-smooth, and let the family interpolate at* x^{int} *in the sense of Definition 1. Write*

$$\Delta_x := f(x) - f^\star, \qquad \Delta_\xi(x) := f_\xi(x) - f_\xi(x^{\mathrm{int}}) \ \geq 0. \tag{45}$$

Then for every $x \in \mathbb{R}^d$,

$$\mathbb{E}_\xi \left\|\nabla f_\xi(x)\right\|^2 \ \leq \ 4L_0\,\Delta_x + 8L_1^2\,\mathbb{E}_\xi\left[\Delta_\xi(x)^2\right]. \tag{46}$$

Assume in addition the bounded dissimilarity *condition*

$$\Delta_\xi(x) \leq \beta\,\Delta_x \qquad \text{for a.e. } \xi \text{ and all } x \in \mathbb{R}^d, \tag{47}$$

for some $\beta \geq 1$. *Then, for every* $\Delta > 0$, *the strong growth condition* (42) *holds on the sublevel set* $\mathcal{S}_\Delta := \{x : \Delta_x \leq \Delta\}$ *with*

$$\begin{aligned} \rho(\Delta) \ &\leq \ \frac{4L_0 + 8L_1^2\beta^2\Delta}{\gamma(\Delta)}, \\ \gamma(\Delta) &:= \inf\left\{\tfrac{\|\nabla f(x)\|^2}{\Delta_x} \ : \ x \in \mathcal{S}_\Delta,\ \Delta_x > 0\right\}. \end{aligned} \tag{48}$$

At points with $\Delta_x = 0$ *one has* $\nabla f_\xi(x) = 0$ *for a.e.* ξ, *so* (42) *holds there trivially for any* $\rho \geq 1$. *If* f *is* μ*-strongly convex then* $\gamma(\Delta) \geq 2\mu$ *and*

$$\rho(\Delta) \leq \frac{2L_0 + 4L_1^2\beta^2\Delta}{\mu} \ \leq \ 2 + \frac{2\beta\big(L_0 + 2L_1^2\beta\Delta\big)}{\mu}. \tag{49}$$

Proof. Fix x and a realization ξ for which (44) holds, and abbreviate $G_\xi := \|\nabla f_\xi(x)\|$, $\Delta_\xi := \Delta_\xi(x)$. Since f_ξ is convex, (L_0, L_1)-smooth and attains its infimum at x^{int}, Lemma 2 applies to f_ξ and gives

$$G_\xi^2 \ \leq \ 2\big(L_0 + L_1 G_\xi\big)\,\Delta_\xi. \tag{50}$$

Viewing (50) as the quadratic inequality $G_\xi^2 - 2L_1\Delta_\xi G_\xi - 2L_0\Delta_\xi \leq 0$ in $G_\xi \geq 0$ and taking the positive root,

$$G_\xi \ \leq \ L_1\Delta_\xi + \sqrt{L_1^2\Delta_\xi^2 + 2L_0\Delta_\xi}. \tag{51}$$

Squaring (51) and using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$G_\xi^2 \ \leq \ 2L_1^2\Delta_\xi^2 + 2\big(L_1^2\Delta_\xi^2 + 2L_0\Delta_\xi\big) \ = \ 4L_1^2\Delta_\xi^2 + 4L_0\Delta_\xi. \tag{52}$$

Taking expectations in (52) and using

$$\mathbb{E}_\xi\big[\Delta_\xi(x)\big] = f(x) - \mathbb{E}_\xi f_\xi(x^{\mathrm{int}}) = f(x) - f^\star = \Delta_x \tag{53}$$

(the middle equality is Definition 1) yields (46), since $4L_1^2 \leq 8L_1^2$.

Now assume (47) and let $x \in \mathcal{S}_\Delta$ with $\Delta_x > 0$. Then $\mathbb{E}_\xi[\Delta_\xi^2] \leq \beta^2\Delta_x^2 \leq \beta^2\Delta\,\Delta_x$, so (46) gives

$$\mathbb{E}_\xi \left\|\nabla f_\xi(x)\right\|^2 \ \leq \ \big(4L_0 + 8L_1^2\beta^2\Delta\big)\,\Delta_x. \tag{54}$$

Dividing (54) by $\|\nabla f(x)\|^2 > 0$ (positive because $\Delta_x > 0$) and bounding $\Delta_x / \|\nabla f(x)\|^2 \leq 1/\gamma(\Delta)$ by the definition of $\gamma(\Delta)$ in (48) proves (42) with the constant (48). If $\Delta_x = 0$ then $x \in X^\star$, hence $f_\xi(x) = f_\xi(x^{\mathrm{int}})$ for a.e. ξ by (44) and (53), and minimality gives $\nabla f_\xi(x) = 0$ a.s., so both sides of (42) vanish. Finally, if f is μ-strongly convex, the standard Polyak–Łojasiewicz inequality $\|\nabla f(x)\|^2 \geq 2\mu\Delta_x$ gives $\gamma(\Delta) \geq 2\mu$, and substituting into (48) yields the first bound in (49); the second follows from $\beta \geq 1$ and $L_0 \geq \mu$. □

Lemma 17 shows that interpolation alone yields only the weak-growth inequality (46), whose second term is *quadratic* in the individual gaps; upgrading it to strong growth genuinely requires extra structure, namely bounded dissimilarity (47) together with a quadratic-growth-type lower bound $\gamma(\Delta) > 0$. Assumption 4 is therefore the primitive hypothesis of this appendix, with interpolation as a motivating special case rather than an equivalent one.

Why multiplicative noise is harmless here. The essential consequence of (42) for algorithm design is that the oracle noise is *multiplicative*: at any query x,

$$\mathbb{E}\,\|g(x,\xi)-\nabla f(x)\|^2 \le \mathbb{E}_\xi\,\|\nabla f_\xi(x)\|^2 \le \rho\,\|\nabla f(x)\|^2\,. \tag{55}$$

Two structural facts of our framework then render this noise harmless. First, gradient scales along the outer loop are certified: Lemma 3 and the cap bookkeeping keep $\|\nabla f\|$ bounded by the running cap, so mini-batches of size $\widetilde{O}(\rho)$ make every certificate (cap updates, terminations) reliable. Second, and less obviously, the noise *floor* of a proximal subproblem with parameter λ is proportional to

$$\rho\,\|\nabla f(\hat{x})\|^2 = \rho\,(\lambda+\eta)^2\,\|\hat{x}-c\|^2\,, \tag{56}$$

where $\eta \ge 0$ is the multiplier of the ball constraint (see (6)); our λ-schedules tie this quantity to the current *certified* optimality gap rather than to the target accuracy, so it contracts together with the iterates.

These two facts make a three-stage *hop-and-solve* variant of ARC-SG (Algorithm 2) converge linearly. After Phase I, well-conditioned proximal levels with $\lambda = 4L_1\hat{G}$ drain large gradients at cost $\widetilde{O}(\rho)$ per level (the *hot* stage); then ball-constrained proximal steps of radius $\frac{1}{2L_1}$ with a gap-proportional λ-schedule "hop" towards $x^\star$, strong convexity converting the certified gap into the distance bound $\|c-x^\star\| \le \sqrt{2\,\mathrm{gap}/\mu}$ that accelerates the hops as the gap shrinks (the *cold* stage); once $x^\star$ is provably inside the current ball, a single run of projected batched SGD with the (L_0,L_1)-stepsize, which tolerates multiplicative noise natively, finishes at a linear rate (the *interior* stage). The price of walking the wide, flat part of the valley is the ε-independent burn-in $\rho\bar{\kappa} + \rho(1+L_0/\mu)^3$; we do not believe the cube is tight.

H.2 Supplementary results in the overparameterized regime

Theorem 3 (Strong growth, strongly convex)**.** *Let Assumptions 1, 2 and 4 hold with the a.s. variant (43), and let f be μ-strongly convex. Then the hop-and-solve variant of ARC-SG (Algorithm 2) with mini-batches of size $\widetilde{O}(\rho)$ at all measurement and certificate steps (Stage-2 cold-stage inner iterations use batches of size $\widetilde{O}\big(\rho(1+L_0/\mu)^{3/2}\big)$ per Lemma 21(c)) returns $\hat{x}$ with $f(\hat{x})-f^\star \le \varepsilon$ with probability at least $1-\alpha$ using*

$$N = \widetilde{O}\Big(\rho\Big(1+\tfrac{L_0}{\mu}\Big)\log_+\big(\tfrac{\Delta_{\mathrm{cert}}}{\varepsilon}\big) \;+\; \rho\,\bar{\kappa} \;+\; \rho\Big(1+\tfrac{L_0}{\mu}\Big)^3\Big) \tag{57}$$

oracle calls; the last two terms are independent of ε, and no σ^2-driven term or noise floor is present. Here $\widetilde{O}$ treats the certificate as an externally supplied input and is monotone in it; under the default $\Delta_{\mathrm{cert}} = \Delta_{\mathrm{def}}$ of Lemma 7 the Stage-0 term is $\widetilde{O}\big(\rho(1+q)^4\big)$ with $q := L_1R_0$ (Proposition 5).

Theorem 4 (Strong growth, convex — drift-aware variant)**.** *Let Assumptions 1, 2 and 4 hold with the a.s. variant (43). Consider the following* drift-aware convex variant *of the hop-and-solve method: Stages 1–2 of Algorithm 2 are run on the regularized objective $f_{\mathrm{reg}} = f + \frac{\mu_{\mathrm{reg}}}{2}\big\|\cdot - c^{\mathrm{I}}\big\|^2$, with $\mu_{\mathrm{reg}} = \varepsilon/\bar{R}^2$, centered at the Stage-0 output c^{I} (a choice essential both for the entry certificate and for the moderate size of the certified smoothness constant below); the oracle is drift-corrected to $\hat{g}_{\mathrm{reg}}(x) = \hat{g}(x) + \mu_{\mathrm{reg}}(x-c^{\mathrm{I}})$; the stages use the strong-convexity constant μ_{reg} and the critical scale $\bar{G}'' := L_0'/L_1$ in place of $\bar{G}$, where $L_0' := L_0 + \mu_{\mathrm{reg}} + \frac{25}{8}L_1\bar{G} = \frac{33}{8}L_0 + \mu_{\mathrm{reg}}$: by the convexity comparison $\|\nabla f(x)\|^2 \le \|\nabla f_{\mathrm{reg}}(x)\|^2 + \big\|\nabla f(c^{\mathrm{I}})\big\|^2$ and the entry bound $\big\|\nabla f(c^{\mathrm{I}})\big\| \le \frac{25}{8}\bar{G}$, the objective f_{reg} is* globally (L_0', L_1)*-smooth in the sense of Assumption 2 (Lemma 23); the Stage-1 transfer step and level target are those of Lemma 20, $s := \frac{1}{400L_1}$ and $\delta_{\mathrm{lvl}} := \lambda s^2/2$; the cap anchors use the batch $B''_{\mathrm{rel}} := \lceil 1.6\cdot 10^5 c_g^2\rho\ell_\star\rceil$ and the update $\hat{G}^+ := \frac{8}{7}\|\hat{g}\| + \frac{\bar{G}''}{28}$ of (67); the Stage-2 exit threshold is $\Delta \le \min\{\mu_{\mathrm{reg}}/(32L_1^2), \varepsilon/2\}$; and Stage 3 is skipped (it is unnecessary, since the fixed target ε is already met at the Stage-2 exit). This variant returns an ε-solution of (1) with probability at least $1-\alpha$ using*

$$N = \widetilde{O}\Big(\rho\,\bar{\kappa} \;+\; \rho\Big(1+\tfrac{L_0\bar{R}^2}{\varepsilon}\Big)^3\Big) \tag{58}$$

oracle calls. This is not *a black-box application of Theorem 3: the almost-sure strong growth of f need not transfer to f_{reg}, because $\nabla f(x)$ and the regularization drift $\mu_{\mathrm{reg}}(x-c^{\mathrm{I}})$ may partly*

Reference	Smoothness	Rate (μ-strongly convex)
Vaswani et al. (2019)	L, SGC(ρ)	$\rho\frac{L}{\mu}$; accelerated: $\sqrt{\rho\frac{L}{\mu}}$
Liu & Belkin (2020)	L, SGC(ρ)	$\sqrt{\rho\frac{L}{\mu}}$
Clipped SGD (Lobanov & Koloskova, 2026)	(L_0, L_1), interpolation	$\rho\frac{L_0}{\mu}$ + burn-in
Theorem 3	(L_0, L_1), SGC-GS	$\rho\frac{L_0}{\mu} + \rho\bar{\kappa} + \rho\frac{L_0^3}{\mu^3}$

Table 4: Strong-growth/interpolation regime: oracle calls in $\widetilde{O}$-notation, with the common $\log_+(\Delta_{\mathrm{cert}}/\varepsilon)$ factor of the leading terms omitted from all entries. SGC(ρ) is the classical strong growth condition; SGC-GS is its generalized-smooth analogue (Assumption 4; our theorems use the a.s. variant (43)); $\bar{\kappa} = 1 + 3L_1R_0$, and the $\rho\bar{\kappa} + \rho L_0^3/\mu^3$ terms are an ε-independent burn-in. Batched implementations of our method use batch size $\widetilde{O}(\rho)$ at measurement and certificate steps (Stage-2 cold-stage inner iterations use $\widetilde{O}\big(\rho(1 + L_0/\mu)^{3/2}\big)$, Lemma 21(c)); single-sample variants of smooth-case accelerated methods achieve the $\sqrt{\rho}$-type dependence shown. Entries reflect the assumptions of the cited works (in-expectation versus high-probability, additive versus multiplicative noise) and are not directly comparable line by line; the printed $\widetilde{O}$-forms treat the certificate Δ_{cert} as an externally supplied input (Proposition 5).

cancel; what survives regularization is the convexity-based comparison $\|\nabla f(x)\|^2 \le \|\nabla f_{\mathrm{reg}}(x)\|^2 + \big\|\nabla f(c^{\mathrm{I}})\big\|^2$*, which yields a* global *smoothness certificate whose first constant is inflated by the center-dependent term* $L_1\big\|\nabla f(c^{\mathrm{I}})\big\|$*, and the centering at* c^{I} *rather than* x^0 *is what keeps that term moderate,* $\big\|\nabla f(c^{\mathrm{I}})\big\| \le \frac{25}{8}\bar{G}$ *(Appendix H.4). With* $A := 1 + L_0\bar{R}^2/\varepsilon$ *the bound reads* $\widetilde{O}(\rho\bar{\kappa} + \rho A^3)$*: the only surviving* $\bar{\kappa}$*-factor is the Stage-0 entry burn-in, which the variant shares with Theorem 3.*

Table 4 positions these bounds against the smooth-case strong-growth literature and the generalized-smooth analysis of Lobanov & Koloskova (2026).

Theorem 3 transfers the interpolation phenomenon of Schmidt & Le Roux (2013); Ma et al. (2018); Vaswani et al. (2019) to generalized smoothness: the ε-dependent term matches the classical smooth-case strong-growth rate $\rho(L/\mu)\log_+(\Delta_{\mathrm{cert}}/\varepsilon)$ with L replaced by L_0; all σ-driven terms vanish; the guarantee is high-probability rather than in-expectation; and arbitrarily large initial gradients cost only an ε-independent burn-in, relative to the generalized-smooth analysis of Lobanov & Koloskova (2026). What our framework does *not* deliver is the accelerated $\sqrt{L_0/\mu}$-dependence available in the smooth setting (Vaswani et al., 2019; Liu & Belkin, 2020): as shown in Appendix H.5, inside any epoch-restarted momentum scheme the strong-growth noise at the wandering extrapolation sequence produces a noise floor $\Theta(\kappa^{3/2})$ times the epoch target, forcing batch sizes that erase the $\sqrt{\kappa}$-gain. We believe this obstruction is an artifact of proof technology rather than of the problem, and record:

Open Problem 2. *Under Assumptions 1, 2 and 4, is there an algorithm with oracle complexity* $\widetilde{O}\big(\mathrm{poly}(\rho)\sqrt{L_0/\mu}\,\log_+(\Delta_{\mathrm{cert}}/\varepsilon)\big)$ *plus an* ε*-independent burn-in for* μ*-strongly convex* f*? Equivalently: can the accelerated stochastic strong-growth analyses of Vaswani et al. (2019); Liu & Belkin (2020) be localized to balls of radius* $O(1/L_1)$*?*

H.3 Standing conventions, relative noise, and Phase I

Throughout the rest of this appendix Assumptions 1, 2, 4 hold with the a.s. variant (43), the convention (25) is in force, and f is μ-strongly convex where stated. Note that μ-strong convexity and (2) force $\mu \le L_0$ (Remark 3). We write

$$\begin{aligned} \nu(x) &:= 2\sqrt{\rho}\,\|\nabla f(x)\|\,, \\ \ell_\star &:= \log\Big(\tfrac{2^{30}(\bar{K}_{\mathrm{hot}}+\bar{K}_{\mathrm{cold}}+m)^3(2+\bar{\kappa})^4(2+L_0/\mu)^4}{\alpha}\Big), \end{aligned} \tag{59}$$

where $\ell_\star$ is an umbrella that dominates (a routine check) every logarithm requested from Proposition 1, Lemma 8 and the measurements below at the confidence levels used; the counters $\bar{K}_{\mathrm{hot}}, \bar{K}_{\mathrm{cold}}, m$ are defined in Algorithm 2.

H.3.1 RELATIVE NOISE AND MEASUREMENTS

Lemma 18 (Relative noise). *Under the a.s. variant, $\zeta(x,\xi) := g(x,\xi) - \nabla f(x)$ satisfies $\|\zeta\| \le (1+\sqrt{\rho})\,\|\nabla f(x)\| \le \nu(x)$ a.s.; in particular ζ is norm-sub-Gaussian with parameter $\nu(x)$, and the average $\bar{\zeta}$ of a batch of size b satisfies $\mathbb{P}[\|\bar{\zeta}\| \ge c_g \nu(x)\sqrt{t/b}] \le 2e^{-t}$ for all $t \ge 1$ (Lemma 8). Consequently a batch of size $B_{\mathrm{rel}} := \lceil 256 c_g^2 \rho\, \ell_\star \rceil$ returns $\hat{g}$ with $\|\hat{g} - \nabla f(x)\| \le \frac{1}{8}\|\nabla f(x)\|$, so $\|\nabla f(x)\| \le \frac{8}{7}\|\hat{g}\|$ and $\|\hat{g}\| \le \frac{9}{8}\|\nabla f(x)\|$, with probability $1 - 2e^{-\ell_\star}$.*

H.3.2 Phase I UNDER INTERPOLATION

Proposition 5. *Run Phase I with $B_1 := \lceil 2500 c_g^2 \rho \Lambda_{\mathrm{lg}}^2 \ell_1 \rceil$, $\ell_1 := \log(16 T_1^{\max}/\alpha)$, all other parameters as in (19). Under the a.s. variant all conclusions of Proposition 2 hold with probability at least $1 - \alpha/4$, at total cost $T_1 B_1 = \widetilde{O}\big(\rho(1 + L_1 R_0)\big)$. Here $\widetilde{O}$ treats the certificate as an externally supplied input and is monotone in it; under the default $\Delta_{\mathrm{cert}} = \Delta_{\mathrm{def}}$ of Lemma 7 one has $\Lambda_{\mathrm{lg}} = \Theta(1+q)$ with $q := L_1 R_0$, hence $B_1 = \widetilde{O}\big(\rho(1+q)^2\big)$ and the true cost is $\widetilde{O}\big(\rho(1+q)^4\big)$ with only polylogarithmic factors in $(q, 1/\alpha)$ hidden—the same two-regime distinction as in Corollary 2, and the $\widetilde{O}$-forms of Theorems 3–4 inherit it.*

Proof. With $\eta := 1/(25\Lambda_{\mathrm{lg}})$, Lemma 18 and a union bound over the $\le T_1^{\max} + 1$ steps give $\|\bar{\zeta}_t\| \le \eta G_t$ at every step. The proof of Proposition 2 interacts with the estimation accuracy only through two inequalities: at every non-terminal step it uses $\|\bar{\zeta}_t\| \le G_t/(23\Lambda_{\mathrm{lg}})$, valid there since $\eta \le 1/(23\Lambda_{\mathrm{lg}})$; at the terminal step it uses the output bound, which now reads $G \le \|\hat{g}\|/(1-\eta) \le 3\bar{G} \cdot \frac{25}{24} \le \frac{25}{8}\bar{G}$ (and, conversely, the run cannot stop while $G > \frac{25}{8}\bar{G}$, since then $\|\hat{g}\| \ge (1-\eta)G > 3\bar{G}$; while it runs, $\|\hat{g}_t\| > 3\bar{G}$ gives $G_t \ge 3\bar{G}/(1+\eta) \ge \frac{23}{8}\bar{G}$, so the supercritical chain is intact). The degenerate branch ($L_1 R_0 \le \frac{1}{24}$) uses only the $t = 0$ measurement and is unaffected.

□

H.3.3 R-STM WITH STATE-DEPENDENT NOISE

Remark 11 (Noise interface). *The proof of Proposition 1 interacts with randomness only through the per-query sub-Gaussian parameter $\varsigma_k = c_b\sigma/\sqrt{b_k}$ of the batched estimates at points of Q_k. Hence, if the single-sample noise is norm-sub-Gaussian with parameter ν_k* uniformly over $x \in Q_k$, *replacing (16) by*

$$b_k := \max\left\{1,\ \left\lceil \frac{C_b \nu_k^2 \log(8KN_{\mathrm{ep}}/\alpha')}{\sqrt{\lambda L_F}\,\bar{\Delta}_k} \right\rceil\right\} \tag{60}$$

(the lower bound $b_k \ge 1$ matters when the state-dependent scale ν_k vanishes, e.g. near the optimum under strong growth) leaves every step of the proof unchanged, and the total statistical cost becomes $\sum_k N_{\mathrm{ep}} b_k \le K N_{\mathrm{ep}} + 24 C_b \ell' \sum_k \nu_k^2/(\lambda\bar{\Delta}_k)$ with $\ell' := \log(8KN_{\mathrm{ep}}/\alpha')$.

Lemma 19 (Gradient scale on epoch balls). *Let a level have center c, cap $\hat{G} \ge \|\nabla f(c)\|$, parameter $\lambda > 0$ and ball radius $r_c \le \frac{1}{2L_1}$, and let $L_F := L_c + \lambda$ with $L_c := 4(L_0 + L_1\hat{G})$ (Lemma 4). During epoch k of R-STM on $F := f + \frac{\lambda}{2}\|\cdot - c\|^2$ over $\mathbb{B}(c, r_c)$, every $x \in Q_k$ satisfies $\|\nabla f(x)\| \le 3L_F\sqrt{2\bar{\Delta}_k/\lambda_F} + \hat{G} + 2\lambda r_c$, where $\lambda_F \ge \lambda$ is the strong convexity modulus used; hence the a.s. noise scale on Q_k obeys $\nu_k^2 \le 12\rho\big(18L_F^2\bar{\Delta}_k/\lambda_F + \hat{G}^2 + 4\lambda^2 r_c^2\big)$.*

Proof. For $x \in Q_k$: $\|\nabla f(x)\| \le \|\nabla F(x)\| + \lambda r_c \le \|\nabla F(x) - \nabla F(x_Q)\| + \|\nabla F(x_Q)\| + \lambda r_c$. By Lemma 3 (which covers the ball-constrained case), $\|\nabla f(x_Q)\| \le \|\nabla f(c)\| \le \hat{G}$, so $\|\nabla F(x_Q)\| \le \hat{G} + \lambda r_c$; and $\|\nabla F(x) - \nabla F(x_Q)\| \le L_F\|x - x_Q\| \le L_F D_k = 3L_F\sqrt{2\bar{\Delta}_k/\lambda_F}$ along segments of $\mathbb{B}(c, r_c)$. Square with $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and multiply by $(2\sqrt{\rho})^2$. □

H.3.4 THE HOP-AND-SOLVE ALGORITHM

Algorithm 2 ARC-SG-interp (interpolation/strong growth; hop-and-solve). Constants: $r_{\text{safe}} = \frac{1}{2L_1}$, $s = \frac{1}{400L_1}$, $B_{\text{rel}} = \lceil 256 c_g^2 \rho \ell_\star \rceil$, $\bar{K}_{\text{hot}} := \lceil 5900 L_0/\mu \rceil + 14$, $\bar{K}_{\text{cold}} := \lceil 108 L_0/\mu \rceil + \lceil 2 \log_2(16(1 + L_0/\mu)) \rceil + 24$, $m := \lceil \log_4(8/\alpha) \rceil$.

Require: x^0, R_0, μ, ε, α, ρ

Stage 0: $c \leftarrow$ Phase I(x^0) with B_1 of Proposition 5; $\hat{G} \leftarrow 4\bar{G}$

while $\hat{G} > 2\bar{G}$ **and** fewer than $\bar{K}_{\text{hot}}$ hot levels used **do**

  Stage 1 (hot): $\lambda \leftarrow 4L_1\hat{G}$;

  $c \leftarrow$ R-STM$\big(f + \frac{\lambda}{2}\|\cdot - c\|^2$ over $\mathbb{B}(c, r_{\text{safe}})$; $H_0 = \frac{\hat{G}}{4L_1}$, $\delta_{\text{lvl}} = \frac{\hat{G}}{80000L_1}$, $\frac{\alpha}{16\bar{K}_{\text{hot}}}\big)$

  $\hat{G} \leftarrow \frac{8}{7}\|$GradEst$(c, B_{\text{rel}})\|$

end while

Stage 2 (cold): $\Delta \leftarrow 2\hat{G}^2/\mu$

for $j = 1, \ldots, \bar{K}_{\text{cold}}$ **while** $\Delta > \frac{\mu}{32L_1^2}$ **do**

  $\theta \leftarrow \min\{\frac{1}{2}, \frac{1}{2L_1}\sqrt{\frac{\mu}{2\Delta}}\}$;

  $\lambda \leftarrow 2L_1^2\theta\Delta$;

  $c \leftarrow$ R-STM$\big(f + \frac{\lambda}{2}\|\cdot - c\|^2$ over $\mathbb{B}(c, r_{\text{safe}})$; $H_0 = \hat{G}r_{\text{safe}}$, $\delta_{\text{lvl}} = \frac{\theta\Delta}{256}$, $\frac{\alpha}{16\bar{K}_{\text{cold}}}\big)$

  $\hat{G} \leftarrow \frac{8}{7}\|$GradEst$(c, B_{\text{rel}})\|$;

  $\Delta \leftarrow (1 - \frac{\theta}{4})\Delta$

end for

Stage 3 (interior): conditional on the Stage-0–2 history, for $i = 1, \ldots, m$ run mutually independent copies (disjoint fresh sample streams, all started from the fixed common point $x_0^{(i)} := c$): T steps of projected batched SGD on f over $\mathbb{B}(c, r_{\text{safe}})$, $x_{t+1} = \Pi(x_t - \frac{1}{4L'}\hat{g}_t)$, batch $\lceil 24\rho \rceil$, $L' := 12(L_0 + L_1\hat{G})$, $T := \max\{0, \lceil \frac{4L'}{\mu} \ln \frac{2L'^2}{L_1^2\varepsilon\mu} \rceil\}$

Two-step selection: measure each output with B_{rel}; discard those with $\|\hat{g}_i\| > 2\sqrt{2L'\varepsilon'}$, $\varepsilon' := \frac{\varepsilon\mu}{16L'}$; re-measure survivors with $B_{\text{sel}} := \lceil 95 c_g^2 \rho \ell_\star \rceil$;

return the survivor with smallest re-measured norm

H.3.5 THE HOT STAGE

Lemma 20 (Hot stage)**.** *On the good event, Stage 1 maintains $\|\nabla f(c)\| \le \hat{G} \le \frac{9}{7}\|\nabla f(c)\|$ after each anchor (the initial $\hat{G} = 4\bar{G}$ is valid by Proposition 5), and: (a) each level has $\lambda = 4L_1\hat{G}$, $L_F \le 10L_1\hat{G}$, condition number $\le \frac{5}{2}$, unconstrained prox ($\|\hat{x} - c\| \le \frac{1}{4L_1}$), and certified $H_0 = \hat{G}^2/\lambda$; (b) with $\Psi := f(c) - f^\star + \frac{\hat{G}}{100L_1}$, every executed level decreases Ψ by at least $\frac{\hat{G}}{600L_1} \ge \frac{\bar{\Delta}}{300}$, where $\bar{\Delta} = L_0/L_1^2$; (c) Stage 1 executes at most $5900L_0/\mu + 14$ levels, exits with $\hat{G} \le 2\bar{G}$, and never triggers the $\bar{K}_{\text{hot}}$ guard; (d) each level costs at most $10^9 C_b\, \rho\, \ell_\star$ oracle calls, so the stage costs $\widetilde{O}(\rho(1 + L_0/\mu))$.*

Proof. (a) While $\hat{G} > 2\bar{G}$ we have $L_0 \le L_1\hat{G}/2$, so $L_c = 4(L_0 + L_1\hat{G}) \le 6L_1\hat{G}$ and $L_F = L_c + \lambda \le 10L_1\hat{G}$, giving $1 + L_F/\lambda \le \frac{7}{2}$; a sharper count, $\kappa = L_F/\lambda \le \frac{10}{4}$, is what we use. Containment: $\|\hat{x} - c\| \le \|\nabla f(c)\|/\lambda \le \frac{1}{4L_1} < r_{\text{safe}}$ by Lemma 3, so Remark 5 certifies H_0, and $\delta_{\text{lvl}} = \lambda s^2/2$ gives $\|c^+ - \hat{x}\| \le s$.

(b) Two cases. *Productive* ($\|\hat{x} - c\| \ge \frac{1}{16L_1}$): $f(\hat{x}) \le f(c) - \frac{\lambda}{2}\|\hat{x} - c\|^2 \le f(c) - \frac{\hat{G}}{128L_1}$; the transfer along $[\hat{x}, c^+] \subset \mathbb{B}(c, 2r_{\text{safe}})$ costs $f(c^+) \le f(\hat{x}) + s\|\nabla f(\hat{x})\| + \frac{L_c}{2}s^2 \le f(\hat{x}) + \frac{\hat{G}}{400L_1} + \frac{\hat{G}}{53000L_1}$, so the net drop is $\ge \frac{\hat{G}}{200L_1}$. The anchor can raise the cap: $\|\nabla f(c^+)\| \le \|\nabla f(\hat{x})\| + L_c s \le \hat{G}(1 + 0.015)$, so $\hat{G}^+ \le \frac{9}{7} \cdot 1.015\,\hat{G} \le 1.31\hat{G}$, and $\Delta\Psi \le -\frac{\hat{G}}{200L_1} + \frac{0.31\hat{G}}{100L_1} \le -\frac{\hat{G}}{600L_1}$. *Collapsing* ($\|\hat{x} - c\| < \frac{1}{16L_1}$): then $\|\nabla f(\hat{x})\| = \lambda\|\hat{x} - c\| < \frac{\hat{G}}{4}$, the objective can increase by at most $s\|\nabla f(\hat{x})\| + \frac{L_c}{2}s^2 \le \frac{\hat{G}}{1550L_1} \le \frac{\hat{G}}{360L_1}$, while $\|\nabla f(c^+)\| \le \frac{\hat{G}}{4} + 0.015\hat{G}$, hence $\hat{G}^+ \le \frac{9}{7} \cdot 0.265\,\hat{G} \le 0.35\hat{G}$ and $\Delta\Psi \le \frac{\hat{G}}{360L_1} - \frac{0.65\hat{G}}{100L_1} \le -\frac{\hat{G}}{300L_1}$. In both cases $\Delta\Psi \le -\frac{\hat{G}}{600L_1} \le -\frac{2\bar{G}}{600L_1} = -\frac{\bar{\Delta}}{300}$.

(c) By μ-strong convexity, $f(c)-f^\star \le \|\nabla f(c)\|\,\|c-x^\star\| \le \|\nabla f(c)\|\sqrt{2(f(c)-f^\star)/\mu}$, so $f(c)-f^\star \le 2\|\nabla f(c)\|^2/\mu$; at entry $\|\nabla f(c_1)\| \le \frac{25}{8}\bar G$, hence $\Psi_0 \le \frac{2\cdot 625}{64}\frac{\bar G^2}{\mu} + \frac{4\bar G}{100L_1} \le 19.6\frac{L_0\bar\Delta}{\mu} + \frac{\bar\Delta}{25}$. Since $\Psi \ge 0$ and each level removes $\ge \bar\Delta/300$, at most $300\Psi_0/\bar\Delta \le 5880L_0/\mu + 12$ levels execute before the loop condition fails, i.e. $\hat G \le 2\bar G$. (This also bounds the caps: $\hat G \le 100L_1\Psi_0$, which the umbrella $\ell_\star$ dominates.)

(d) By (a), $H_0/\delta_{\mathrm{lvl}} = 20000$, so $K \le 15$ epochs of $N_{\mathrm{ep}} \le 24\sqrt{5/2}+1 \le 39$ iterations. Remark 11 with Lemma 19 ($r_c = r_{\mathrm{safe}}$, $\lambda_F = \lambda$) gives per-level statistical cost at most $24C_b\ell_\star\big[216\rho(\frac{L_F}{\lambda})^{3/2}KN_{\mathrm{ep}}/N_{\mathrm{ep}} + \dots\big]$; carrying out the three pieces as in Appendix E.2.5: the $L_F^2\bar\Delta_k/\lambda$-piece contributes $\le 5184C_b\rho\ell_\star\kappa^2K \le 5\cdot 10^5 C_b\rho\ell_\star$; the $\hat G^2$-piece contributes $\le 1152C_b\rho\ell_\star\hat G^2/(\lambda\delta_{\mathrm{lvl}}) = 1152\cdot 2\cdot 10^4 C_b\rho\ell_\star$ (since $\lambda\delta_{\mathrm{lvl}} = \hat G^2/20000$); the $\lambda^2 r_{\mathrm{safe}}^2$-piece contributes $\le 4608C_b\rho\ell_\star\lambda r_{\mathrm{safe}}^2/\delta_{\mathrm{lvl}} = 4608\cdot 80000\, C_b\rho\ell_\star$ (since $\lambda r_{\mathrm{safe}}^2/\delta_{\mathrm{lvl}} = 2r_{\mathrm{safe}}^2/s^2 = 80000$). Adding $25K\sqrt{5/2}$ deterministic-equivalent iterations and B_{rel} gives the claim.

□

H.3.6 THE COLD STAGE

Lemma 21 (Cold hops)**.** *On the good event, Stage 2 satisfies, with Δ_j the running schedule value and θ_j, λ_j as in Algorithm 2: (a) $f(c_j)-f^\star \le \Delta_j$ for every hop j, and $\|c_j - x^\star\| \le \sqrt{2\Delta_j/\mu}$; (b) the stage exits within $108L_0/\mu + 2\log_2(16(1+L_0/\mu)) + 22$ hops with $\Delta \le \frac{\mu}{32L_1^2}$, hence $\|c-x^\star\| \le \frac{r_{\mathrm{safe}}}{2}$; the $\bar K_{\mathrm{cold}}$ guard never triggers; (c) the total cost of the stage is $\widetilde O\big(\rho(1+L_0/\mu)^3\big)$.*

Proof. (a) Induction; the base is $f(c_1)-f^\star \le 2\|\nabla f(c_1)\|^2/\mu \le 2\hat G^2/\mu = \Delta_1$ as in Lemma 20(c). Given the invariant, $D_j := \|c_j - x^\star\| \le \sqrt{2\Delta_j/\mu}$. Let $\tilde x$ be the ball-constrained prox. If $D_j \le r_{\mathrm{safe}}$ then $f(\tilde x)-f^\star \le F(x^\star)-f^\star = \frac{\lambda_j}{2}D_j^2 \le \frac{\lambda_j r_{\mathrm{safe}}^2}{2} = \frac{\theta_j\Delta_j}{4}$ (the last identity is the design $\lambda_j = 2L_1^2\theta_j\Delta_j$). Otherwise set $\theta' := r_{\mathrm{safe}}/D_j \in [\theta_j, 1]$ (using $D_j \le \sqrt{2\Delta_j/\mu}$) and $u := c_j + \theta'(x^\star - c_j)$: $f(\tilde x)+\frac{\lambda_j}{2}\|\tilde x - c_j\|^2 \le f(u)+\frac{\lambda_j}{2}r_{\mathrm{safe}}^2$ and convexity give $f(\tilde x)-f^\star \le (1-\theta')\big(f(c_j)-f^\star\big)+\frac{\theta_j\Delta_j}{4} \le (1-\theta_j)\Delta_j + \frac{\theta_j\Delta_j}{4}$. R-STM (Remark 11) returns c_{j+1} with $F(c_{j+1})-F(\tilde x) \le \delta_j = \frac{\theta_j\Delta_j}{256}$ and $d := \|c_{j+1}-\tilde x\| \le \sqrt{2\delta_j/\lambda_j} = \frac{1}{16L_1}$. Then, by the two-case argument of (33) with $a := \|\tilde x - c_j\| \le r_{\mathrm{safe}}$, $f(c_{j+1}) \le F(c_{j+1}) - \frac{\lambda_j}{2}\|c_{j+1}-c_j\|^2 \le f(\tilde x) + \delta_j + \lambda_j\, a\, d \le f(\tilde x) + \delta_j + \lambda_j r_{\mathrm{safe}}\, d$, and $\lambda_j r_{\mathrm{safe}}\, d = 2L_1^2\theta_j\Delta_j\cdot\frac{1}{2L_1}\cdot\frac{1}{16L_1} = \frac{\theta_j\Delta_j}{16}$. Collecting, $f(c_{j+1})-f^\star \le (1-\theta_j)\Delta_j + \theta_j\Delta_j(\frac14 + \frac{1}{16}+\frac{1}{256}) \le (1-\frac{\theta_j}{4})\Delta_j = \Delta_{j+1}$.

(b) While $\theta_j = \frac12$ (i.e. $\Delta_j \le \frac{\mu}{2L_1^2}$) the schedule contracts by $\frac78$ per hop, so at most $\lceil\log_{8/7}16\rceil = 21$ such hops occur before $\Delta \le \frac{\mu}{32L_1^2}$. In the regime $\theta_j < \frac12$, on a dyadic band $\Delta \in (\Delta', 2\Delta']$ one has $\theta_j \ge \frac{1}{2L_1}\sqrt{\mu/(4\Delta')}$, so at most $\lceil 4\ln 2/\theta\rceil \le (16\ln 2)\,L_1\sqrt{\Delta'/\mu}+1$ hops halve the schedule. Summing the geometric series over bands below $\Delta_1 \le 8\bar G^2/\mu$ gives at most $38L_1\sqrt{\Delta_1/\mu} + \log_2\frac{\Delta_1}{\Delta_{\mathrm{stop}}} \le 108\frac{L_0}{\mu} + \log_2\big(256\frac{L_0^2}{\mu^2}\big)$ hops. At exit $\Delta \le \frac{\mu r_{\mathrm{safe}}^2}{8}$, so $\|c-x^\star\| \le \sqrt{2\Delta/\mu} \le \frac{r_{\mathrm{safe}}}{2}$ by (a).

(c) By (a) and Lemma 2, $\|\nabla f(c_j)\| \le 2L_1\Delta_j + \sqrt{2L_0\Delta_j}$ regardless of how c_j was produced, so the anchored cap obeys $\hat G_j \le \frac97\|\nabla f(c_j)\| \le 3L_1\Delta_j + 2\sqrt{L_0\Delta_j}$ and $L_c \le 4L_0 + 4L_1\hat G_j \le 8L_0 + 16L_1^2\Delta_j$ (AM–GM), $L_F \le 8L_0 + 17L_1^2\Delta_j$. In the regime $\theta_j < \frac12$, $\lambda_j = L_1\sqrt{\mu\Delta_j/2}$ and $\kappa_j := L_F/\lambda_j \le \frac{12L_0}{L_1\sqrt{\mu\Delta_j}} + 25L_1\sqrt{\Delta_j/\mu}$; over the admissible range $\Delta_j \in (\frac{\mu}{32L_1^2}, 8\bar G^2/\mu]$ both terms are at most $25\sqrt8\,\frac{L_0}{\mu}$ each, and in the regime $\theta_j = \frac12$ ($\lambda_j = L_1^2\Delta_j$), $\kappa_j \le 17 + \frac{8L_0}{L_1^2\Delta_j} \le 17 + 256\frac{L_0}{\mu}$; in all cases $\kappa_j \le 300(1+\frac{L_0}{\mu})$. Per hop, Remark 11 with Lemma 19 gives cost $25K_j\sqrt{\kappa_j} + 5184C_b\rho\ell_\star\kappa_j^2K_j + 1152C_b\rho\ell_\star\frac{\hat G_j^2}{\lambda_j\delta_j} + 4608C_b\rho\ell_\star\frac{\lambda_j r_{\mathrm{safe}}^2}{\delta_j} + B_{\mathrm{rel}}$, with $K_j \le \ell_\star$. The last ratio equals $\frac{\lambda_j r_{\mathrm{safe}}^2}{\delta_j} = \frac{\theta_j\Delta_j/2}{\theta_j\Delta_j/256} = 128$, an absolute constant. For the

$\hat{G}^2$-piece, in the regime $\theta_j < \frac{1}{2}$ one has $\lambda_j\delta_j = \frac{\mu\Delta_j}{1024}$ and $\hat{G}_j^2 \le 18L_1^2\Delta_j^2 + 8L_0\Delta_j$, so it is at most $1.2\cdot 10^6 C_b\rho\ell_\star(18L_1^2\Delta_j + 8L_0)/\mu$; summed over hops with the band counts of (b), the L_0-part gives $\widetilde{O}(\rho(L_0/\mu)^2)$ and the $L_1^2\Delta_j$-part, dominated by the top band $\Delta_1 \le 8\bar{G}^2/\mu$ through $\sum_j \Delta_j \le 34\Delta_1^{3/2}L_1/\sqrt{\mu}$, gives $\widetilde{O}(\rho(L_0/\mu)^3)$; in the regime $\theta_j = \frac{1}{2}$ the piece is $\widetilde{O}(\rho(1+L_0/\mu))$ per hop. For the κ_j^2-piece, the same band accounting yields $\sum_j \kappa_j^2 = \widetilde{O}\big((L_0/\mu)^3 + (1+L_0/\mu)^2\big)$: the square of the second term of κ_j, $625L_1^2\Delta_j/\mu$, summed against band counts $(16\ln 2)\,L_1\sqrt{\Delta_j/\mu}$ is again dominated by Δ_1 and gives the cube; the square of the first term, $144L_0^2/(L_1^2\mu\Delta_j)$, summed against the same counts is dominated by Δ_{stop} and gives $\widetilde{O}((L_0/\mu)^2)$. Multiplying by $5184C_b\rho\ell_\star K_j$ and adding the deterministic part $\sum_j 25\ell_\star\sqrt{\kappa_j} = \widetilde{O}((1+L_0/\mu)^{3/2})$ proves (c). (The cubes originate in the early, wide-valley hops, where the certified gradient scale can be as large as $L_1\Delta_j$; we do not believe they are tight, cf. Remark 14.) □

H.3.7 THE INTERIOR STAGE

Remark 12 (Extended ball smoothness and co-coercivity)**.** *The proof of Lemma 4 applies on $\mathbb{B}(c,\,4r)$ for $r \le \frac{1}{2L_1}$ with the constant $L' := 12(L_0 + L_1\hat{G})$, using the sharp form of Lemma 6 ($G_z \le e^{L_1 s}G_c + \frac{L_0}{L_1}(e^{L_1 s}-1)$ with $s = \|z-c\| \le 4r$): the Grönwall factor becomes $e^{4L_1 r} \le e^2$, so $L_0 + L_1 G_z \le e^2(L_0 + L_1\hat{G})$ on the ball, and the two-point bound sums to $\|\nabla f(x) - \nabla f(y)\| \le e^2(L_0 + L_1\hat{G})\,\|x-y\| \le L'\,\|x-y\|$, since $e^2 \le 12$. Consequently, if $x^\star \in \mathbb{B}(c,\,r_{\mathrm{safe}})$, then for every $x \in \mathbb{B}(c,\,r_{\mathrm{safe}})$ the virtual point $y := x - \nabla f(x)/L'$ stays in the certified region: $\|\nabla f(x)\| \le \hat{G} + L' r_{\mathrm{safe}}$ gives $\|y - x\| \le \frac{\hat{G}}{L'} + r_{\mathrm{safe}} \le \frac{1}{12L_1} + \frac{1}{2L_1}$, so $y \in \mathbb{B}\Big(c,\,\frac{13}{12L_1}\Big) \subset \mathbb{B}(c,\,4r_{\mathrm{safe}})$, and the descent lemma along $[x,y]$ yields the co-coercivity bound $\|\nabla f(x)\|^2 \le 2L'\big(f(x) - f^\star\big)$ on $\mathbb{B}(c,\,r_{\mathrm{safe}})$.*

Lemma 22 (Interior SGD and selection)**.** *Suppose Stage 3 starts with $\|c - x^\star\| \le \frac{r_{\mathrm{safe}}}{2}$ and a valid cap $\hat{G} \ge \|\nabla f(c)\|$. Then, conditionally on the Stage-0–2 history: (a) each SGD run satisfies $\mathbb{E}[\|x_T - x^\star\|^2 \mid \mathcal{F}_2] \le (1 - \frac{\mu}{4L'})^T\,\|c - x^\star\|^2$ and, with the stated T, $\mathbb{E}[f(x_T) - f^\star \mid \mathcal{F}_2] \le \varepsilon'/4$; (b) with conditional probability at least $1 - \alpha/4$, the two-step selection returns $\hat{x}$ with $f(\hat{x}) - f^\star \le \frac{\varepsilon}{4}$. Consequently the same probability bound holds unconditionally. (c) The stage costs $\widetilde{O}\big(\rho(1 + \frac{L_0}{\mu})\log_+(\frac{L_0^2}{L_1^2\varepsilon\mu})\cdot\log_+(1/\alpha)\big)$ oracle calls. Moreover at entry $\hat{G} \le \bar{G}$, so $L' \le 24L_0$, and $\mathcal{F}_2$ denotes the history through Stage 2.*

Proof. Fix the Stage-0–2 history throughout; all expectations and probabilities below are conditional on $\mathcal{F}_2$ until the final unconditioning step. At entry, Lemma 2 with $f(c) - f^\star \le \frac{\mu}{32L_1^2}$ and $\mu \le L_0$ gives $\|\nabla f(c)\| \le \frac{\mu}{16L_1} + \sqrt{\frac{L_0\mu}{16L_1^2}} \le \frac{\bar{G}}{16} + \frac{\bar{G}}{4}$, so the anchored cap is $\le \frac{9}{7}\cdot\frac{5}{16}\bar{G} \le \bar{G}$ and $L' \le 24L_0$. *(a)* Projection onto $\mathbb{B}(c,\,r_{\mathrm{safe}})$ is nonexpansive and $x^\star$ is feasible, so with $r_t := \|x_t - x^\star\|$, $\mathbb{E}r_{t+1}^2 \le r_t^2 - 2\gamma\,\langle\nabla f(x_t),\,x_t - x^\star\rangle + \gamma^2\mathbb{E}\,\|\hat{g}_t\|^2$. Strong convexity gives $\langle\nabla f,\,x - x^\star\rangle \ge f(x) - f^\star + \frac{\mu}{2}r^2$; the batch of size $\ge 24\rho$ gives $\mathbb{E}\,\|\bar{\zeta}\|^2 \le \frac{4\rho}{24\rho}\,\|\nabla f\|^2 \le \frac{1}{6}\,\|\nabla f\|^2$, so $\mathbb{E}\,\|\hat{g}\|^2 \le \frac{7}{6}\,\|\nabla f\|^2 \le \frac{7}{3}L'(f - f^\star)$ by Remark 12. With $\gamma = \frac{1}{4L'}$, $\mathbb{E}r_{t+1}^2 \le (1 - \gamma\mu)r_t^2 - 2\gamma(1 - \frac{7}{24})(f - f^\star) \le (1 - \frac{\mu}{4L'})r_t^2$. Since $[x^\star, x_T] \subset \mathbb{B}(c,\,r_{\mathrm{safe}})$ and f is L'-smooth there, $\mathbb{E}[f(x_T) - f^\star] \le \frac{L'}{2}\mathbb{E}r_T^2 \le \frac{L'}{2}e^{-T\mu/(4L')}\frac{1}{16L_1^2} \le \frac{\varepsilon'}{4}$ by the choice of T (if the logarithm is non-positive then $T = 0$, and the entry certificate $f(c) - f^\star \le \frac{\mu}{32L_1^2} \le \frac{\varepsilon'}{4}$ already delivers the target). *(b)* By Markov and (a), each run fails ($f(x_T^{(i)}) - f^\star > \varepsilon'$) with probability $\le \frac{1}{4}$. The m runs use disjoint, mutually independent sample batches and share only the history-measurable starting point c and the parameters (L', T) computed before the stage. Conditionally on the Stage-0–2 history these quantities are fixed and the failure events are independent, so the probability that all runs fail simultaneously is $\le 4^{-m} \le \frac{\alpha}{8}$ by the choice $m = \lceil\log_4(8/\alpha)\rceil$. A good run has $G_i := \Big\|\nabla f(x_T^{(i)})\Big\| \le \sqrt{2L'\varepsilon'}$ (Remark 12), so its relative measurement is $\le \frac{9}{8}\sqrt{2L'\varepsilon'} < 2\sqrt{2L'\varepsilon'}$ and it survives; every survivor has $G_i \le \frac{8}{7}\cdot 2\sqrt{2L'\varepsilon'} = \frac{16}{7}\sqrt{2L'\varepsilon'}$, hence noise scale $\nu_i \le \frac{32}{7}\sqrt{2\rho L'\varepsilon'}$, and the batch $B_{\mathrm{sel}} =$

$\lceil 95c_g^2\rho\ell_\star\rceil$ yields absolute accuracy $c_g\nu_i\sqrt{\ell_\star/B_{\mathrm{sel}}} \le \frac{\sqrt{\mu\varepsilon}}{6}$ (using $L'\varepsilon' = \frac{\varepsilon\mu}{16}$). If $\hat{x}$ minimizes the re-measured norm among survivors, then $\|\nabla f(\hat{x})\| \le \sqrt{2L'\varepsilon'} + 2\cdot\frac{\sqrt{\mu\varepsilon}}{6} = \sqrt{\frac{\mu\varepsilon}{8}} + \frac{\sqrt{\mu\varepsilon}}{3}$, and $f(\hat{x}) - f^\star \le \frac{\|\nabla f(\hat{x})\|^2}{2\mu} \le \frac{(0.354+0.334)^2}{2}\varepsilon \le \frac{\varepsilon}{4}$. A union bound over the $\le 2m$ measurements, each failing with probability at most $2e^{-\ell_\star}$, adds at most $4me^{-\ell_\star} \le \alpha/2^{28}$ (by the definition (59) of $\ell_\star$). Hence the total Stage-3 failure probability is at most $\alpha/8 + \alpha/2^{28} < \alpha/4$, exactly as stated. *(c)* Immediate: $m\big(T\lceil 24\rho\rceil + B_{\mathrm{rel}} + B_{\mathrm{sel}}\big)$. □

H.4 Proofs of Theorems 3 and 4

H.4.1 Proof of Theorem 3

On the intersection of the good events of Proposition 5 and Lemmas 20–22, the output satisfies $f(\hat{x}) - f^\star \le \frac{\varepsilon}{4}$. Here is the explicit confidence accounting: Stage 0 consumes at most $\alpha/4$; the hot and cold R-STM calls consume at most $\alpha/16$ each; all hot/cold cap-anchor measurements together consume less than $\alpha/16$ by (59); and Lemma 22 consumes less than $\alpha/4$. Thus the total failure probability is at most $11\alpha/16 < \alpha$. The cost is the sum of: Stage 0, $\widetilde{O}(\rho(1 + L_1R_0)) = \widetilde{O}(\rho\bar{\kappa})$; Stage 1, $\widetilde{O}(\rho(1+L_0/\mu))$; Stage 2, $\widetilde{O}(\rho(1+L_0/\mu)^3)$; Stage 3, $\widetilde{O}(\rho(1+L_0/\mu)\log_+(\Delta_{\mathrm{cert}}/\varepsilon))$ after absorbing $\log_+(L_0/\mu)$ and $\log_+(1/\alpha)$ factors into $\widetilde{O}$.

□

H.4.2 Proof of Theorem 4

The regularization must be centered at the Phase I output. A remark on the construction is in order, because the naive choice is not merely suboptimal but leaves a genuine gap. Regularizing around the *initial* point, $f + \frac{\mu_{\mathrm{reg}}}{2}\left\|\cdot - x^0\right\|^2$, breaks the entry certificate of the hot stage: Stage 0 is run on f and Proposition 5 certifies a cap for $\|\nabla f(c)\|$ only, whereas the hot stage of Algorithm 2 must be started from a valid cap for

$$\|\nabla f_{\mathrm{reg}}(c)\| = \left\|\nabla f(c) + \mu_{\mathrm{reg}}(c - x^0)\right\|, \tag{61}$$

and the drift term $\mu_{\mathrm{reg}}\left\|c - x^0\right\|$ in (61) is not controlled by any *a priori* bound of order $\bar{G}''$; bounding it through the hot-stage potential would be circular, since that potential is itself established only after the entry cap is known to be valid. The same choice would also defeat the smoothness certificate of Lemma 23 below, whose first constant is inflated by $L_1\|\nabla f(c)\|$ at the center c: at $c = x^0$ this is the arbitrarily large $L_1\left\|\nabla f(x^0)\right\|$, while at $c = c^{\mathrm{I}}$ it is at most $\frac{25}{8}L_1\bar{G} = \frac{25}{8}L_0$. We therefore center the regularizer at the Phase I output.

Construction. Let c^{I} be the point returned by Stage 0 (i.e., Phase I run on f; Proposition 5), and set

$$\begin{aligned} f_{\mathrm{reg}} &:= f + \tfrac{\mu_{\mathrm{reg}}}{2}\left\|\cdot - c^{\mathrm{I}}\right\|^2, \qquad \mu_{\mathrm{reg}} := \frac{\varepsilon}{\bar{R}^2}, \\ L_0' &:= L_0 + \mu_{\mathrm{reg}} + \tfrac{25}{8}L_1\bar{G} \;=\; \tfrac{33}{8}L_0 + \mu_{\mathrm{reg}}, \qquad \bar{G}'' := \frac{L_0'}{L_1}, \end{aligned} \tag{62}$$

and write $x^\star_{\mathrm{reg}} := \arg\min f_{\mathrm{reg}}$, $\bar{\Delta}' := L_0'/L_1^2$, $A := 1 + L_0\bar{R}^2/\varepsilon$ and $A' := L_0'/\mu_{\mathrm{reg}} = 1 + \frac{33}{8}L_0\bar{R}^2/\varepsilon$, so that $A \le A' \le \frac{33}{8}A$. Then f_{reg} is μ_{reg}-strongly convex. The *naive* global pair $(L_0 + \mu_{\mathrm{reg}}, L_1)$ for f_{reg}, however, need not satisfy Assumption 2: $\nabla f(x)$ and the drift $\mu_{\mathrm{reg}}(x - c^{\mathrm{I}})$ may partly cancel, in which case $\|\nabla f_{\mathrm{reg}}(x)\|$ is small while the local Lipschitz constant of ∇f_{reg} retains the scale $L_1\|\nabla f(x)\| + \mu_{\mathrm{reg}}$, of order $L_1\mu_{\mathrm{reg}}\left\|x - c^{\mathrm{I}}\right\|$, which the naive first constant does not dominate. Concretely, $f(x) = \cosh x$ satisfies $|f''| \le 1 + |f'|$ (hence Assumption 2 with the pair (e, e) via Remark 4), but $f_{\mathrm{reg}}(x) = \cosh x + \frac{1}{2}(x - 100)^2$ has, at its own minimizer $\bar{x}$, $\nabla f_{\mathrm{reg}}(\bar{x}) = 0$ and $f''_{\mathrm{reg}}(\bar{x}) = \cosh\bar{x} + 1 \approx 95.76 > e+1$; (2) for f_{reg} with the naive pair $(e+1, e)$, applied at $x = \bar{x}$, would force $|f'_{\mathrm{reg}}(\bar{x}+t) - f'_{\mathrm{reg}}(\bar{x})| \le (e+1)\,|t|$ for $|t| \le 1/e$, i.e. $f''_{\mathrm{reg}}(\bar{x}) \le e + 1$—a contradiction. What fails is only the naive first constant: Lemma 23 below shows that f_{reg} always admits a *global* pair with the same second constant L_1, whose first constant is inflated by $L_1\left\|\nabla f(c^{\mathrm{I}})\right\|$ (in the example the center is $c = 100$ and the inflation is the enormous $L_1\sinh 100$, which illustrates

precisely why the center's gradient must be certified small); at the actual center, Proposition 5 gives $\|\nabla f(c^{\mathrm{I}})\| \le \frac{25}{8}\bar{G}$, producing L_0' of (62). Concretely, Stages 1–2 of Algorithm 2 are run on f_{reg}, using the corrected oracle

$$\hat{g}_{\mathrm{reg}}(x) := \hat{g}(x) + \mu_{\mathrm{reg}}(x - c^{\mathrm{I}}) \tag{63}$$

(the correction is deterministic and free of oracle calls), with $\bar{G}$ replaced by $\bar{G}''$ in all stage thresholds, with the Stage-1 transfer step and level target as in Lemma 20, $s := \frac{1}{400L_1}$ and $\delta_{\mathrm{lvl}} := \lambda s^2/2$, with cap anchors of batch $B''_{\mathrm{rel}} := \lceil 1.6 \cdot 10^5 c_g^2 \rho \ell_\star \rceil$ and update (67), with the guards $\bar{K}_{\mathrm{hot}} := \lceil 2950A' \rceil + 8$ and $\bar{K}_{\mathrm{cold}} := \lceil 108A' \rceil + \lceil 2\log_2(16(1+A')) \rceil + 24$ in place of the corresponding counters of Algorithm 2 (the cold counter is that of Algorithm 2 with $\mu \mapsto \mu_{\mathrm{reg}}$, $L_0/\mu \mapsto A'$; the hot one reflects the sharper anchor tracking of Step 3), with the logarithmic umbrella $\ell_\star$ of (59) evaluated at the variant's counters ($\mu \mapsto \mu_{\mathrm{reg}}$, $L_0 \mapsto L_0'$), and with the Stage-2 exit threshold replaced by $\Delta \le \min\{\frac{\mu_{\mathrm{reg}}}{32L_1^2}, \frac{\varepsilon}{2}\}$; Stage 3 is skipped.

Lemma 23 (Global regularized certificate)**.** *Let f be convex, differentiable and (L_0, L_1)-smooth in the sense of Assumption 2, let $c \in \mathbb{R}^d$ and $\mu > 0$, and set $F := f + \frac{\mu}{2}\|\cdot - c\|^2$ and $G_c := \|\nabla f(c)\|$. Then, for every $x \in \mathbb{R}^d$,*

$$\|\nabla f(x)\|^2 \;\le\; \|\nabla F(x)\|^2 + G_c^2, \qquad \|\nabla f(x)\| \;\le\; \|\nabla F(x)\| + G_c, \tag{64}$$

and F satisfies Assumption 2 globally *with the pair $(L_0 + \mu + L_1 G_c,\ L_1)$, on the same admissible range $\|y - x\| \le 1/L_1$. The squared comparison is sharp: equality holds for affine f at the minimizer of F.*

Proof. Write $r := \|x - c\|$. By gradient monotonicity, $\langle \nabla f(x), x - c\rangle \ge \langle \nabla f(c), x - c\rangle \ge -G_c\, r$, so expanding $\|\nabla F(x)\|^2 = \|\nabla f(x)\|^2 + 2\mu\,\langle \nabla f(x), x - c\rangle + \mu^2 r^2$ gives

$$\|\nabla F(x)\|^2 \;\ge\; \|\nabla f(x)\|^2 + \mu r\,(\mu r - 2G_c).$$

If $\mu r \ge 2G_c$ the last term is nonnegative and $\|\nabla f(x)\| \le \|\nabla F(x)\|$; if $\mu r < 2G_c$ then $\mu r\,(2G_c - \mu r) \le G_c^2$ (the left side is maximized at $\mu r = G_c$), so $\|\nabla f(x)\|^2 \le \|\nabla F(x)\|^2 + G_c^2$. This proves the first inequality of (64), and the second follows from $\sqrt{a^2 + b^2} \le a + b$; for affine $f(x) = \langle a, x\rangle$ the minimizer $x = c - a/\mu$ of F satisfies $\nabla F(x) = 0$ and $\|\nabla f(x)\|^2 = G_c^2$, so the squared comparison holds with equality. (The plain triangle inequality $\|\nabla f(x)\| \le \|\nabla F(x)\| + \mu r$, with the worst uncompensated case $\mu r \approx 2G_c$ of the case split, yields only $\|\nabla f(x)\| \le \|\nabla F(x)\| + 2G_c$ and hence the looser first constant $L_0 + \mu + 2L_1 G_c$; instantiated at $c = c^{\mathrm{I}}$ this reads $\frac{29}{4}L_0 + \mu_{\mathrm{reg}}$ in place of $\frac{33}{8}L_0 + \mu_{\mathrm{reg}}$. We use the sharp comparison.) Finally, for $\|y - x\| \le 1/L_1$, (2) applied to f together with (64) gives

$$\begin{aligned}\|\nabla F(y) - \nabla F(x)\| &\le \big(L_0 + L_1\,\|\nabla f(x)\|\,\big)\,\|y - x\| + \mu\,\|y - x\| \\ &\le \big(L_0 + \mu + L_1\,G_c + L_1\,\|\nabla F(x)\|\,\big)\,\|y - x\|\,,\end{aligned}$$

which is (2) for F with the stated pair. □

Step 1: the entry cap is exact. Since $\nabla f_{\mathrm{reg}}(c^{\mathrm{I}}) = \nabla f(c^{\mathrm{I}})$, the cap produced by Stage 0 transfers verbatim:

$$\|\nabla f_{\mathrm{reg}}(c^{\mathrm{I}})\| = \|\nabla f(c^{\mathrm{I}})\| \le \tfrac{25}{8}\bar{G} \le \tfrac{25}{8}\bar{G}'' \le 4\bar{G}'', \tag{65}$$

so $\hat{G}_1 := 4\bar{G}''$ is valid, exactly as required by Lemma 20, with no measurement and no appeal to any potential. This is the step that fails for the x^0-centered regularizer.

Step 2: global smoothness and the regularized noise. Instantiating Lemma 23 at $(c, \mu) = (c^{\mathrm{I}}, \mu_{\mathrm{reg}})$, where $G_{c^{\mathrm{I}}} = \|\nabla f(c^{\mathrm{I}})\| \le \frac{25}{8}\bar{G}$ by Proposition 5, certifies f_{reg} *globally* with the pair (L_0', L_1) of (62). Two consequences are the only smoothness inputs that Stages 1–2 use: (i) Lemma 4 applies to f_{reg} on every ball $\mathbb{B}(c, 2r)$ with $r \le \frac{1}{2L_1}$, giving two-point L_c-smoothness with $L_c := 4(L_0' + L_1\hat{G})$ whenever $\hat{G} \ge \|\nabla f_{\mathrm{reg}}(c)\|$, and the descent inequality with modulus L_c along every segment of such a ball—exactly the hypothesis of Proposition 1 and Lemma 19; and (ii) Lemma 2 applies to f_{reg} at every point, with $L_0 \mapsto L_0'$. Hence Lemmas 20 and 21 apply to f_{reg} with the

substitution $(L_0, \mu, \bar{G}) \mapsto (L_0', \mu_{\mathrm{reg}}, \bar{G}'')$, and no level-dependent constant, displacement bound, or induction over levels is needed.

The oracle noise of f_{reg} is that of f: the drift correction in (63) is deterministic and cancels in $\zeta_{\mathrm{reg}}(x,\xi) := \hat{g}_{\mathrm{reg}}(x,\xi) - \nabla f_{\mathrm{reg}}(x) = \hat{g}(x,\xi) - \nabla f(x)$, so, by Lemma 18 and (64),

$$\begin{aligned} \|\zeta_{\mathrm{reg}}\| &\le 2\sqrt{\rho}\,\|\nabla f(x)\| \\ &\le 2\sqrt{\rho}\,\|\nabla f_{\mathrm{reg}}(x)\| + 2\sqrt{\rho}\,G_{c^{\mathrm{I}}} \\ &\le 2\sqrt{\rho}\,\|\nabla f_{\mathrm{reg}}(x)\| + \sigma_{\mathrm{add}}, \\ \sigma_{\mathrm{add}} &:= \tfrac{25}{4}\sqrt{\rho}\,\bar{G} \le \tfrac{25}{4}\sqrt{\rho}\,\bar{G}''. \end{aligned} \tag{66}$$

Every query of Stages 1–2 therefore sees norm-sub-Gaussian noise that is *multiplicative* in $\|\nabla f_{\mathrm{reg}}\|$—the structure that Lemma 19 and Remark 11 already accommodate—plus one additive scale σ_{add} that is constant across levels; its cost is accounted for in Steps 3 and 5.

Step 3: consequences for the analysis. *Cap anchors.* By (66), the noise at the anchor point c_{j+1} has scale $\nu(c_{j+1}) \le 2\sqrt{\rho}\,\|\nabla f_{\mathrm{reg}}(c_{j+1})\| + \sigma_{\mathrm{add}}$ at every level, hot or cold, regardless of how large $\hat{G}_j$ is relative to $\bar{G}''$. With batch $B''_{\mathrm{rel}} = \lceil 1.6 \cdot 10^5 c_g^2 \rho \ell_\star \rceil$, for which $c_g\sqrt{\rho\ell_\star / B''_{\mathrm{rel}}} \le 1/400$, the estimate therefore satisfies

$$\begin{aligned} \|\hat{g} - \nabla f_{\mathrm{reg}}(c_{j+1})\| &\le c_g\,\nu(c_{j+1})\sqrt{\tfrac{\ell_\star}{B''_{\mathrm{rel}}}} \le \frac{2\,\|\nabla f_{\mathrm{reg}}(c_{j+1})\|}{400} + \frac{25\,\bar{G}''}{4\cdot 400} \\ &= \frac{\|\nabla f_{\mathrm{reg}}(c_{j+1})\|}{200} + \frac{\bar{G}''}{64} \le \tfrac{1}{8}\,\|\nabla f_{\mathrm{reg}}(c_{j+1})\| + \tfrac{1}{32}\bar{G}'', \end{aligned}$$

so the update

$$\hat{G}^+ := \tfrac{8}{7}\,\|\hat{g}\| + \tfrac{\bar{G}''}{28} \tag{67}$$

is valid and tracks the true gradient: on the one hand $\hat{G}^+ \ge \frac{8}{7}\big(\frac{199}{200}\,\|\nabla f_{\mathrm{reg}}\| - \frac{\bar{G}''}{64}\big) + \frac{\bar{G}''}{28} \ge \|\nabla f_{\mathrm{reg}}\|$, since $\frac{8}{7}\cdot\frac{199}{200} \ge 1$ and $\frac{\bar{G}''}{28} \ge \frac{8}{7}\cdot\frac{\bar{G}''}{64} = \frac{\bar{G}''}{56}$; on the other hand $\hat{G}^+ \le \frac{8}{7}\big(\frac{201}{200}\,\|\nabla f_{\mathrm{reg}}\| + \frac{\bar{G}''}{64}\big) + \frac{\bar{G}''}{28} \le \frac{9}{7}\,\|\nabla f_{\mathrm{reg}}\| + \frac{\bar{G}''}{14}$.

Hot stage. Lemma 20 applies to f_{reg} with $(L_0, \mu, \bar{G}) \mapsto (L_0', \mu_{\mathrm{reg}}, \bar{G}'')$: while $\hat{G} > 2\bar{G}''$ one has $L_0' = L_1\bar{G}'' \le L_1\hat{G}/2$, so $L_c = 4(L_0' + L_1\hat{G}) \le 6L_1\hat{G}$, $L_F = L_c + \lambda \le 10L_1\hat{G}$ and $\kappa = L_F/\lambda \le \frac{5}{2}$—exactly Lemma 20(a). The prox remains unconstrained: $\|\hat{x} - c\| = \|\nabla f_{\mathrm{reg}}(\hat{x})\|/\lambda \le \hat{G}/\lambda = \frac{1}{4L_1} < r_{\mathrm{safe}}$ by Lemma 3, so $H_0 = \hat{G}^2/\lambda$ is certified as in Remark 5, and $\delta_{\mathrm{lvl}} = \lambda s^2/2$ gives $\|c^+ - \hat{x}\| \le s$. The naive cap control of Lemma 20(b) suffices as stated: $\|\nabla f_{\mathrm{reg}}(c^+)\| \le \|\nabla f_{\mathrm{reg}}(\hat{x})\| + L_c s$ with $L_c s \le 6L_1\hat{G}\cdot\frac{1}{400L_1} = 0.015\,\hat{G}$. The only change to the arithmetic of Lemma 20(b) is the anchor slack $\frac{3}{56}\bar{G}'' \le \frac{3}{112}\hat{G}$ in the tracking bound of (67): with $\Psi := f_{\mathrm{reg}}(c) - f^\star_{\mathrm{reg}} + \frac{\hat{G}}{100L_1}$, a productive level ($\|\hat{x} - c\| \ge \frac{1}{16L_1}$) drops the objective by $\frac{\hat{G}}{128L_1} - \frac{\hat{G}}{400L_1} - \frac{\hat{G}}{53000L_1} \ge \frac{\hat{G}}{189L_1}$ (prox decrease minus transfer along $[\hat{x}, c^+]$, as in Lemma 20(b)) and raises the cap to $\hat{G}^+ \le \frac{8}{7}\cdot\frac{201}{200}\cdot 1.015\,\hat{G} + \frac{3}{112}\hat{G} \le 1.193\,\hat{G}$, so $\Delta\Psi \le -\frac{\hat{G}}{189L_1} + \frac{0.193\,\hat{G}}{100L_1} \le -\frac{\hat{G}}{300L_1}$; a collapsing level raises the objective by at most $\frac{\hat{G}}{1600L_1} + \frac{\hat{G}}{53000L_1} \le \frac{\hat{G}}{1550L_1}$ while $\hat{G}^+ \le \frac{8}{7}\cdot\frac{201}{200}\cdot 0.265\,\hat{G} + \frac{3}{112}\hat{G} \le 0.332\,\hat{G}$, so $\Delta\Psi \le \frac{\hat{G}}{1550L_1} - \frac{0.668\,\hat{G}}{100L_1} \le -\frac{\hat{G}}{166L_1}$. In both cases

$$\begin{aligned} \Delta\Psi \;&\le\; -\frac{\hat{G}}{300L_1} \;\le\; -\frac{2\bar{G}''}{300L_1} \;=\; -\frac{\bar{\Delta}'}{150}, \qquad \text{while} \\ \Psi_0 &\le \frac{2\,\|\nabla f_{\mathrm{reg}}(c_1)\|^2}{\mu_{\mathrm{reg}}} + \frac{4\bar{G}''}{100L_1} \le \frac{625}{32}\,\frac{\bar{G}''^2}{\mu_{\mathrm{reg}}} + \frac{\bar{\Delta}'}{25} \le 19.6\,\bar{\Delta}'A' + \frac{\bar{\Delta}'}{25} \end{aligned} \tag{68}$$

(using $\bar{G}''^2/\mu_{\mathrm{reg}} = \bar{\Delta}'\cdot L_0'/\mu_{\mathrm{reg}} = \bar{\Delta}'A'$ and (65)), so Stage 1 executes at most $150\,\Psi_0/\bar{\Delta}' \le 2940A' + 6$ levels, exits with $\hat{G} \le 2\bar{G}''$, and the guard $\bar{K}_{\mathrm{hot}}$ never triggers. Per level, the cost is that of Lemma 20(d), $\widetilde{O}(\rho)$, plus the σ_{add}-piece $O\big(C_b\ell_\star\sigma_{\mathrm{add}}^2/(\lambda\delta_{\mathrm{lvl}})\big) = O(\rho)$, since $\lambda\delta_{\mathrm{lvl}} = \lambda^2 s^2/2 = \hat{G}^2/20000$ and $\sigma_{\mathrm{add}}^2 = O(\rho\bar{G}''^2)$ with $\hat{G} > 2\bar{G}''$; the anchor batch $B''_{\mathrm{rel}} = O(\rho\ell_\star)$ is of the same order. Hence each hot level costs $\widetilde{O}(\rho)$, and the hot stage costs $\widetilde{O}(\rho A')$ in total.

Cold stage. Lemma 21(a)–(b) goes through for f_{reg} verbatim with $(L_0, \mu, \bar{G}) \mapsto (L_0', \mu_{\mathrm{reg}}, \bar{G}'')$: the argument uses only convexity, μ_{reg}-strong convexity, the (smoothness-free) transfer inequality (33), the R-STM guarantee, and the schedule, which is unchanged; the hop count of Lemma 21(b), a schedule fact, becomes $108A' + 2\log_2(16(1+A')) + 22$ via $\Delta_1 \le 8\bar{G}''^2/\mu_{\mathrm{reg}}$ and $\Delta_{\mathrm{stop}} = \mu_{\mathrm{reg}}/(32L_1^2)$, and the guard $\bar{K}_{\mathrm{cold}}$ never triggers. The gradient-dependent constants follow from the now *global* localization of Step 2, $\|\nabla f_{\mathrm{reg}}(c_j)\| \le 2L_1\Delta_j + \sqrt{2L_0'\Delta_j}$, and the anchor tracking of (67):

$$\hat{G}_j \;\le\; \tfrac{9}{7}\,\|\nabla f_{\mathrm{reg}}(c_j)\| + \tfrac{1}{14}\bar{G}'' \;\le\; \tfrac{9}{7}\big(2L_1\Delta_j + \sqrt{2L_0'\Delta_j}\big) + \tfrac{1}{14}\bar{G}'' \;=\; O\big(L_1\Delta_j + \sqrt{L_0'\Delta_j} + \bar{G}''\big), \tag{69}$$

hence $L_{c,j} \le 4L_0' + 4L_1\hat{G}_j = O\big(L_0' + L_1^2\Delta_j + L_1\sqrt{L_0'\Delta_j}\,\big)$ (using $L_1\bar{G}'' = L_0'$), and the condition-number computation of Lemma 21(c) with $(L_0, \mu) \mapsto (L_0', \mu_{\mathrm{reg}})$ gives $\kappa_j := L_{F,j}/\lambda_j = O(A')$ in both cold regimes.

Step 4: exit threshold and output accuracy. The cold loop exits with $\Delta \le \min\{\frac{\mu_{\mathrm{reg}}}{32L_1^2}, \frac{\varepsilon}{2}\}$, and by (25), $32L_1^2\bar{R}^2 \ge 2$, hence $\frac{\mu_{\mathrm{reg}}}{32L_1^2} = \frac{\varepsilon}{32L_1^2\bar{R}^2} \le \frac{\varepsilon}{2}$: the two thresholds coincide, the loop is exactly that of Algorithm 2 with the substitution of Step 2, and no boundary accounting of the hop count is needed. At the exit the certified gap is $f_{\mathrm{reg}}(\hat{x}) - f_{\mathrm{reg}}^\star \le \frac{\varepsilon}{2}$. To transfer it to f, note that optimality of $x_{\mathrm{reg}}^\star$ for f_{reg} against the feasible point $x^\star$ gives $\frac{\mu_{\mathrm{reg}}}{2}\left\|x_{\mathrm{reg}}^\star - c^{\mathrm{I}}\right\|^2 \le f(x^\star) - f(x_{\mathrm{reg}}^\star) + \frac{\mu_{\mathrm{reg}}}{2}\left\|x^\star - c^{\mathrm{I}}\right\|^2 \le \frac{\mu_{\mathrm{reg}}}{2}\left\|x^\star - c^{\mathrm{I}}\right\|^2$, so, with Proposition 2(iii),

$$\left\|x_{\mathrm{reg}}^\star - c^{\mathrm{I}}\right\| \;\le\; \left\|x^\star - c^{\mathrm{I}}\right\| \;\le\; \tfrac{3}{2}R_0 = \tfrac{\bar{R}}{2}. \tag{70}$$

Therefore

$$f(\hat{x}) - f^\star \;\le\; \big(f_{\mathrm{reg}}(\hat{x}) - f_{\mathrm{reg}}(x_{\mathrm{reg}}^\star)\big) + \tfrac{\mu_{\mathrm{reg}}}{2}\left\|x^\star - c^{\mathrm{I}}\right\|^2 \le\; \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2\bar{R}^2}\cdot\tfrac{\bar{R}^2}{4} \;=\; \tfrac{5\varepsilon}{8} \;\le\; \varepsilon. \tag{71}$$

Step 5: cost. Stage 0 costs $\widetilde{O}(\rho(1 + L_1R_0)) = \widetilde{O}(\rho\bar{\kappa})$ (Proposition 5), and the hot stage costs $\widetilde{O}(\rho A')$ (Step 3). For the cold stage, the multiplicative pieces are those of Lemma 21(c) with $(L_0, \mu) \mapsto (L_0', \mu_{\mathrm{reg}})$ and sum to $\widetilde{O}(\rho A'^3)$, with two modifications. First, the cap (69) carries the anchor slack: $\hat{G}_j^2 = O(L_1^2\Delta_j^2 + L_0'\Delta_j + \bar{G}''^2)$ against $\lambda_j\delta_j = \frac{\mu_{\mathrm{reg}}\Delta_j}{1024}$ (regime $\theta_j < \frac{1}{2}$); the $L_1^2\Delta_j$-part is dominated by the top band $\Delta_1 \le 8\bar{G}''^2/\mu_{\mathrm{reg}}$ and gives $\widetilde{O}(\rho A'^3)$, the L_0'-part is $O(\rho A')$ per hop and totals $\widetilde{O}(\rho A'^2)$ over the $O(A')$ hops, and the $\bar{G}''^2/\Delta_j$-part is dominated by the bottom band. The bottom band is $\Delta_{\mathrm{stop}} = \frac{\mu_{\mathrm{reg}}}{32L_1^2}$—not $\frac{\varepsilon}{2}$: under (25) one has $\Delta_{\mathrm{stop}} \le \frac{\varepsilon}{2}$ (Step 4), and the distinction affects only absolute constants, since $L_1\bar{G}'' = L_0'$ collapses it to the factor $\sqrt{32}$:

$$\widetilde{O}\Big(\rho\,\bar{G}''^2\,\frac{L_1}{\mu_{\mathrm{reg}}^{3/2}\sqrt{\Delta_{\mathrm{stop}}}}\Big) \;=\; \widetilde{O}\Big(\rho\,\bar{G}''^2\,\frac{\sqrt{32}\,L_1^2}{\mu_{\mathrm{reg}}^2}\Big) \;=\; \widetilde{O}\Big(\rho\,\Big(\frac{L_1\bar{G}''}{\mu_{\mathrm{reg}}}\Big)^2\Big) \;=\; \widetilde{O}(\rho A'^2), \tag{72}$$

where the first expression is the per-hop cost $O\big(\rho\bar{G}''^2\ell_\star/(\mu_{\mathrm{reg}}\Delta_j)\big)$ multiplied by the per-band hop count $(16\ln 2)\,L_1\sqrt{\Delta'/\mu_{\mathrm{reg}}} + 1$ of Lemma 21(b) and summed over the dyadic bands (the sum is dominated by the bottom band; the $+1$ of the band count contributes a geometric series over the bands, dominated by the bottom band—at most twice the largest per-hop term—and is equally absorbed by $\widetilde{O}$). Second, the additive-noise piece: per hop it costs $O\big(C_b\ell_\star\sigma_{\mathrm{add}}^2/(\lambda_j\delta_j)\big)$, which has exactly the $\bar{G}''^2$-form just summed, since $\sigma_{\mathrm{add}}^2 = O(\rho\bar{G}''^2)$ is constant across levels; it therefore contributes the same $\widetilde{O}(\rho A'^2)$ in the regime $\theta_j < \frac{1}{2}$, and in the regime $\theta_j = \frac{1}{2}$, where $\lambda_j\delta_j = L_1^2\Delta_j^2/512$, the geometric sum of $1/\Delta_j^2$ over $\Delta_j \ge \Delta_{\mathrm{stop}}$ is dominated by $\Delta_j = \Delta_{\mathrm{stop}}$, and $\bar{G}''^2/(L_1^2\Delta_{\mathrm{stop}}^2) = 1024\,(L_1\bar{G}''/\mu_{\mathrm{reg}})^2 = O(A'^2)$ over $O(1)$ such hops. The deterministic part $\sum_j 25K_j\sqrt{\kappa_j} = \widetilde{O}(A'^{3/2})$ and the anchor batches $O(A')\cdot B_{\mathrm{rel}}'' = \widetilde{O}(\rho A')$ are dominated by the terms above. The cold stage therefore costs $\widetilde{O}\big(\rho A'^3 + \rho A'^2\big)$. Adding the three groups and using $A \le A' \le \frac{33}{8}A$ gives

$$N = \widetilde{O}\big(\rho\,\bar{\kappa} + \rho\,A' + \rho\,A'^2 + \rho\,A'^3\big) \;=\; \widetilde{O}\big(\rho\,\bar{\kappa} + \rho\,A'^3\big) = \widetilde{O}\Big(\rho\,\bar{\kappa} + \rho\Big(1 + \tfrac{L_0\bar{R}^2}{\varepsilon}\Big)^3\Big), \tag{73}$$

which is the statement of Theorem 4: the only surviving $\bar{\kappa}$-factor is the Stage-0 entry burn-in, shared with Theorem 3, and under the default certificate of Lemma 7 that term is $\widetilde{O}\big(\rho(1+q)^4\big)$ with $q := L_1 R_0$, exactly as in Theorem 3 (Proposition 5).

□

H.4.3 TOWARDS IN-EXPECTATION STRONG GROWTH

Remark 13 (Towards in-expectation strong growth)**.** *If only* (42) *holds (no a.s. bound), single-sample noise need not have sub-Gaussian tails, and our proofs do* not *extend by a drop-in replacement of batch means. Two distinct uses of estimates must be separated. (i)* Measurements *(cap anchors, the two-step selection): these only require a high-probability tail bound on a point estimate of a fixed vector, and a geometric median-of-means (MoM) estimator (Minsker, 2015) over $O(\log(1/\alpha'))$ groups delivers sub-Gaussian-type confidence from the second-moment bound $\mathbb{E}\,\|g-\nabla f\|^2 \le \rho\,\|\nabla f\|^2$; all measurement steps of Algorithm 2 therefore survive under* (42) *alone, at the cost of absolute constants. (ii)* Inner-solver gradients*: here the obstruction is real. Median-of-means is a* biased *estimator of the gradient, while the analysis of Proposition 1 decomposes the error as a martingale-difference term (Lemma 9) plus a squared-norm term, and the martingale part relies on conditional unbiasedness of $\hat{g}_t$ given the past; a per-iteration MoM estimate breaks this centering, and controlling the resulting bias along an accelerated trajectory would require exactly the kind of high-relative-accuracy estimates whose avoidance is the point of the design. Plain batch means, on the other hand, have only the second-moment bound under* (42)*, which yields in-expectation but not high-probability control of* (15)*. We therefore state our results under the a.s. variant and leave the in-expectation extension of Theorems 3–4 under* (42) *alone as an open technical question; we expect a restart-and-select outer wrapper around in-expectation inner runs to work, but a complete argument must handle the bias–variance interplay at the extrapolation sequence and we do not claim it here.*

H.5 THE $\kappa^{3/2}$ OBSTRUCTION BEHIND OPEN PROBLEM 2

Remark 14 (Noise floor inside momentum epochs)**.** *Suppose one replaces the interior SGD by an accelerated inner solver, e.g. R-STM on f itself over $\mathbb{B}(c, r_{\mathrm{safe}})$ with $\lambda \asymp \mu$ and $\kappa := L'/\mu$. By Lemma 19 with $\hat{G} \asymp 0$ at the optimum, the noise scale on epoch k is $\nu_k^2 \asymp \rho L'^2 \bar{\Delta}_k/\mu$, so the batch rule of Remark 11 forces $b_k \asymp C_b \rho \ell \,\frac{L'^2/\mu}{\sqrt{\mu L'}} = C_b \rho \ell\, \kappa^{3/2}$* per iteration*: inside a momentum epoch the extrapolation sequence wanders a distance $\Theta(\sqrt{\bar{\Delta}_k/\mu})$ from the optimum, where the strong-growth noise is $\Theta(\sqrt{\rho}\,L'\sqrt{\bar{\Delta}_k/\mu})$—a factor $\Theta(\sqrt{\kappa})$ larger than the noise SGD sees at comparable suboptimality—and killing it inside condition* (16) *costs the $\kappa^{3/2}$ batch. The resulting complexity, $\sqrt{\kappa}\,\log_+(\Delta_{\mathrm{cert}}/\varepsilon)\cdot \rho\kappa^{3/2} = \rho\kappa^2 \log_+(\Delta_{\mathrm{cert}}/\varepsilon)$, is worse than the $\rho\kappa\log_+(\Delta_{\mathrm{cert}}/\varepsilon)$ of plain SGD: the $\sqrt{\kappa}$ gain of momentum is exactly erased. Any resolution must either certify the extrapolation sequence at $o(\sqrt{\kappa})$ relative noise or avoid epoch-uniform batching; we leave this as Open Problem 2.*

I COMPLETE EXPERIMENTAL DETAILS

This section gives the complete experimental protocol, diagnostics, ablations, sensitivity analysis, projection and wall-clock accounting, and the schedule-faithful sanity experiment. The two principal curves remain in Figure 2 of the main paper and are not reproduced here.

Instances. To obtain *global* (L_0, L_1)-smoothness with known constants we use objectives that are separable over rotated coordinates $y = Q^\top x$ for an orthogonal $Q = [q_1, \ldots, q_d]$, with a random root vector r; then the Hessian is diagonal in the rotated basis and it suffices to certify each coordinate term g_k against its own derivative, since $|g_k'(y_k)| \le \|\nabla f(x)\|$. Two term types are used. *Cosh terms*

$\sum_s \cosh(C_{k,s}(y_k - r_k))$: within a coordinate all residuals share a sign, so, using $\cosh t \le 1 + |\sinh t|$,

$$\begin{aligned} g_k''(y_k) &= \sum_s C_{k,s}^2 \cosh(\cdot) \\ &\le \sum_s C_{k,s}^2 + \big(\max_s C_{k,s}\big) \Big| \sum_s C_{k,s} \sinh(\cdot) \Big| \\ &= \sum_s C_{k,s}^2 + \big(\max_s C_{k,s}\big)\, |g_k'(y_k)|. \end{aligned}$$

Quartic terms $g_k(u) = (c\,u)^4$: from $g_k'' = 12c^4u^2$ and $g_k' = 4c^4u^3$, splitting at $|cu| = 1$ gives $g_k'' \le 12c^2 + 3c\,|g_k'|$, i.e. each quartic coordinate is exactly $(12c^2, 3c)$-generalized smooth; its Hessian *vanishes* at the root. These are *Hessian-form* constants only. Assumption 2, however, is the finite-difference condition (2) on the range $\|y - x\| \le 1/L_1$, and the Hessian-form pairs do *not* satisfy it: on the cosh champion coordinate ($C_s \in \{0.6, 0.8, 1.0\}$, pair $(2, 1)$, admissible range $|y - x| \le 1$),

$$\begin{aligned} |g'(1) - g'(0)| &= 0.6\sinh(0.6) + 0.8\sinh(0.8) + \sinh(1) \\ &= 2.2677 > 2 = \big(L_0 + L_1|g'(0)|\big) \cdot |1 - 0|, \end{aligned}$$

and on an individual term $20\cosh(u)$ of the overparameterized family of Appendix J.1 (pair $(20, 1)$), $|g'(1) - g'(0)| = 20\sinh(1) = 23.504 > 20$. The exact finite-difference frontier is as follows.

Lemma 24 (Certified pairs for the experimental families)**.** *(i) Let the finite, nonempty family of scales satisfy $C_s \ge 0$ and set $B := \max_s C_s$. If $B = 0$, the 1-D sum $g(u) = \sum_s \cosh(C_s u)$ is constant and satisfies (2) for every $L_0 > 0$, $L_1 \ge 0$. If $B > 0$, no pair with $L_1 = 0$ is admissible; for $L_1 > 0$, the sum satisfies (2) if and only if $L_1 \ge B/\ln 2$ and $L_0 \ge L_1 \sum_s C_s \sinh(C_s/L_1)$. (Equivalently, arbitrary real scales may be replaced by $|C_s|$, since $\cosh(C_s u) = \cosh(|C_s|u)$.) (ii) $g(u) = (cu)^4$ satisfies (2) if and only if $L_0 L_1^2 \ge (24 + 16\sqrt{2})\, c^4$. (iii) If every coordinate term of $f(x) = \sum_k g_k((Q^\top x)_k)$ with orthogonal Q satisfies (2) with the same pair (L_0, L_1), then so does f.*

Proof. (i) If $B = 0$, then g is constant and the claim is immediate, so assume $B > 0$. Any admissible pair must then have $L_1 > 0$, because $L_1 = 0$ would require the nonconstant cosh sum to have a globally Lipschitz derivative, whereas its second derivative is unbounded. *Necessity.* Taking $x = 0$, $y = 1/L_1$ in (2) and using $g'(0) = 0$ gives $|g'(y) - g'(x)| = g'(1/L_1) = \sum_s C_s \sinh(C_s/L_1)$, while $(L_0 + L_1|g'(x)|)\,|y - x| = L_0/L_1$, so $L_0 \ge L_1 \sum_s C_s \sinh(C_s/L_1)$. For the second condition take $y = x + 1/L_1$ and let $x \to +\infty$: by the addition formula, $\sinh(C_s y) - \sinh(C_s x) = \frac{1}{2} e^{C_s x}\big(e^{C_s/L_1} - 1\big) + O(e^{-C_s x})$. Writing $M \ge 1$ for the multiplicity of the maximal scale $C_s = B$, the maximal-scale terms dominate both displays and give

$$|g'(y) - g'(x)| = \tfrac{1}{2} M B e^{Bx}\big(e^{B/L_1} - 1\big) + o(e^{Bx}), \qquad L_1|g'(x)| = \tfrac{1}{2} L_1 M B e^{Bx} + o(e^{Bx}),$$

while $|y - x| = 1/L_1$ and L_0 is negligible against e^{Bx}. The ratio $|g'(y) - g'(x)| / [(L_0 + L_1|g'(x)|)\,|y - x|]$ therefore tends to $e^{B/L_1} - 1$, forcing $e^{B/L_1} - 1 \le 1$, i.e. $L_1 \ge B/\ln 2$.

Sufficiency. Step 1 (sign sharing). Every term $C_s \sinh(C_s u)$ has the sign of u, so $|g'(u)| = \sum_s C_s |\sinh(C_s u)|$. *Step 2 (a scalar inequality).* For all real X, τ,

$$|\sinh(X + \tau) - \sinh X| \;\le\; \sinh|\tau| + \big(e^{|\tau|} - 1\big)\,|\sinh X|. \tag{74}$$

Both sides are invariant under $(X, \tau) \mapsto (-X, -\tau)$, so assume $X \ge 0$. *Case $\tau \ge 0$:* expanding $\sinh(X + \tau) = \sinh X \cosh\tau + \cosh X \sinh\tau$ and using $e^\tau - \cosh\tau = \sinh\tau$ and $\cosh X - \sinh X = e^{-X}$,

$$\text{RHS} - \text{LHS} \;=\; \sinh\tau\,(1 - \cosh X) + \sinh X\,(e^\tau - \cosh\tau) \;=\; \sinh\tau\,\big(1 - e^{-X}\big) \;\ge\; 0.$$

Case $\tau = -t$, $0 \le t \le X$: the left side is $\sinh X - \sinh(X - t) = 2\cosh(X - \frac{t}{2})\sinh\frac{t}{2} \le \cosh X \sinh t$ (since $0 \le X - \frac{t}{2} \le X$ gives $\cosh(X - \frac{t}{2}) \le \cosh X$, and $2\sinh\frac{t}{2} \le 2\sinh\frac{t}{2}\cosh\frac{t}{2} = \sinh t$), and $\cosh X \sinh t \le \sinh t + (e^t - 1)\sinh X$ because $\cosh X - 1 \le \sinh X$ (equivalently $e^{-X} \le 1$) and $\sinh t \le e^t - 1$ (equivalently $\cosh t \ge 1$). *Case $\tau = -t$, $t > X$:* then $|\sinh(X - t) - \sinh X| = \sinh(t - X) + \sinh X \le \sinh\big((t - X) + X\big) = \sinh t$ by the superadditivity of sinh on $[0, \infty)$ (sinh is convex with $\sinh 0 = 0$, so $\sinh(a + b) \ge \sinh a + \sinh b$ for

$a, b \geq 0$), and $\sinh t$ is absorbed by the right side of (74). *Step 3 (sub-bounds for $|h| \leq 1/L_1$, $C \leq B$).* Since $s \mapsto \sinh s/s$ is increasing on $s > 0$ and $C|h| \leq C/L_1$, $\sinh(C|h|) \leq L_1|h|\sinh(C/L_1)$; and since $t \mapsto e^t - 1$ is convex with value 0 at the origin, $e^{C|h|} - 1 \leq L_1|h|\left(e^{C/L_1} - 1\right) \leq L_1|h|$, using $e^{C/L_1} \leq e^{B/L_1} \leq 2$. *Step 4 (assembly).* Applying (74) termwise with $X = C_s u$, $\tau = C_s h$, summing, and using Steps 1 and 3,

$$\begin{aligned} |g'(u+h) - g'(u)| \;&\leq\; \sum_s C_s \Big[\sinh(C_s|h|) + \left(e^{C_s|h|} - 1\right) |\sinh(C_s u)| \Big] \\ &\leq\; L_1|h| \sum_s C_s \Big[\sinh(C_s/L_1) + |\sinh(C_s u)| \Big] \\ &=\; L_1|h| \Big[\sum_s C_s \sinh(C_s/L_1) + |g'(u)| \Big] \;\leq\; \left(L_0 + L_1|g'(u)|\right)|h|, \end{aligned}$$

by the choice of L_0; this is (2) at the pair $(x, y) = (u, u+h)$, with $|h| \leq 1/L_1$ arbitrary.

(ii) If $c = 0$, then g is constant and the claim is immediate. Assume $c \neq 0$. At $L_1 = 0$ the derivative of this quartic is not globally Lipschitz, and the displayed algebraic condition also fails, so it remains to consider $L_1 > 0$. Write $y = x + h$. Since $(x+h)^3 - x^3 = h\,q(h)$ with $q(h) := 3x^2 + 3xh + h^2$, and $q(h) \geq 0$ for all real h (its discriminant is $9x^2 - 12x^2 = -3x^2 \leq 0$), one has $|g'(y) - g'(x)| = 4c^4|h|\,q(h)$, while $|g'(x)| = 4c^4|x|^3$. The quadratic q is convex in h, so its maximum over $|h| \leq 1/L_1$ is attained at an endpoint; for $x \geq 0$ the maximizer is $h = +1/L_1$, and by the symmetry $(x, h) \mapsto (-x, -h)$ the binding requirement is

$$L_0 \;\geq\; 4c^4 \max_{x \geq 0} \Phi(x), \qquad \Phi(x) := 3x^2 + \tfrac{3x}{L_1} + \tfrac{1}{L_1^2} - L_1 x^3.$$

Now $\Phi'(x) = 6x + \frac{3}{L_1} - 3L_1x^2$ vanishes at $x^\star = \frac{1+\sqrt{2}}{L_1}$ (the positive root of $L_1x^2 - 2x - \frac{1}{L_1} = 0$), which is the global maximizer on $[0, \infty)$: $\Phi''(x^\star) = -6\sqrt{2} < 0$, the value $\Phi(0) = \frac{1}{L_1^2}$ is smaller, and $\Phi(x) \to -\infty$ as $x \to \infty$. With $t := L_1 x^\star = 1 + \sqrt{2}$ (so $t^2 = 3 + 2\sqrt{2}$ and $t^3 = 7 + 5\sqrt{2}$),

$$L_1^2 \Phi(x^\star) \;=\; 3t^2 + 3t + 1 - t^3 \;=\; (9 + 6\sqrt{2}) + (3 + 3\sqrt{2}) + 1 - (7 + 5\sqrt{2}) \;=\; 6 + 4\sqrt{2},$$

giving the requirement $L_0 L_1^2 \geq 4(6 + 4\sqrt{2})\,c^4 = (24 + 16\sqrt{2})\,c^4$. Necessity is attained by the witness $(x, y) = (x^\star,\, x^\star + 1/L_1)$, for which equality holds in every step above.

(iii) Let $u = Q^\top x$, $v = Q^\top y$; since Q is orthogonal, $|v_k - u_k| \leq \|v - u\| = \|y - x\| \leq 1/L_1$, so the per-coordinate assumption applies to each pair (u_k, v_k), and

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\|^2 \;&=\; \sum_k |g_k'(v_k) - g_k'(u_k)|^2 \;\leq\; \sum_k \left(L_0 + L_1|g_k'(u_k)|\right)^2 |v_k - u_k|^2 \\ &\leq\; \max_k \left(L_0 + L_1|g_k'(u_k)|\right)^2 \|v - u\|^2 \;\leq\; \left(L_0 + L_1 \|\nabla f(x)\|\right)^2 \|y - x\|^2, \end{aligned}$$

using $|g_k'(u_k)| \leq \|\nabla f(x)\|$ (each coordinate of $g'(u)$ is, up to the orthogonal rotation, a coordinate of $\nabla f(x)$) and the monotonicity of $t \mapsto (L_0 + L_1 t)^2$ on $t \geq 0$. □

Applied to the champion scales $\{0.6, 0.8, 1.0\}$ ($B = 1$), part (i) gives the tight pair

$$(L_0, L_1) = \Big(\tfrac{1}{\ln 2} \big[0.6 \sinh(0.6 \ln 2) + 0.8 \sinh(0.8 \ln 2) + \sinh(\ln 2)\big],\; \tfrac{1}{\ln 2} \Big) = (2.1258,\; 1.4427),$$

with admissible radius $1/L_1 = \ln 2 \approx 0.693$; the required L_1 scales as $c_k/\ln 2$ and the required L_0 is increasing in the base scale $c_k \leq 1$, so the same pair certifies every $\cosh$ coordinate, and for the quartic coordinates $L_0 L_1^2 \approx 4.42 \geq (24 + 16\sqrt{2})/81 \approx 0.5756$, so part (ii) is non-binding (even their Hessian-form pair $(\frac{4}{3}, 1)$ clears the frontier). By part (iii) the pair lifts to the full objectives. The generic conversion of Remark 4 would certify only the looser pairs $(2e, e)$ or $(2/\ln 2, 1/\ln 2)$; the tight pairs above are used in all reported runs. (a) *Convex:* $d = 40$; coordinates 1–30 are cosh with base scales c_k geometrically spaced in $[0.35, 1]$ times inner scales $\{0.6, 0.8, 1.0\}$ (90 components, the frontier driven by the $c_k = 1$ coordinate), coordinates 31–40 are quartic with $c = \frac{1}{3}$. Hence $L_0 = 2.1258$, $L_1 = 1/\ln 2 \approx 1.4427$, both tight for the champion coordinate; $x^\star = Qr$ and $f^\star = 90$ (the number of cosh components) exactly; f is convex with a unique minimizer but not strongly

convex, globally or locally at $x^\star$. The start is $x^0 = x^\star + t\, q_{k_\star}$ along a quartic coordinate, with t found by bisection so that $\|\nabla f(x^0)\| = 150\,\bar{G}$ ($R_0 \approx 16.5$). (b) *Strongly convex:* $d = 40$, all coordinates cosh with three scales (120 components); since $g_k'' \geq \sum_s C_{k,s}^2$ pointwise, f is globally μ-strongly convex with the exact constant $\mu = \min_k \sum_s C_{k,s}^2 = 2 \cdot 0.35^2 = 0.245$, with the same certified pair $(L_0, L_1) = (2.1258, 1/\ln 2)$ as in (a); no quadratic term is added, so $\|\nabla^2 f(x^\star)\| \leq L_0$ and all constants supplied to the methods are simultaneously valid; $f^\star = 120$ exactly. The interpolating family used in Appendix J.1 is built the same way and described there.

Oracle, noise calibration, budgets, statistics. In (a)–(b) the oracle returns the full gradient corrupted by additive Gaussian noise, $\nabla f(x) + \frac{s_g}{\sqrt{b}}\,\xi$ with $\xi \sim \mathcal{N}(0, I_d)$ for a batch of size b and generation scale $s_g = 0.25$. Assumption 3 requires $\mathbb{E}\exp(\|\zeta\|^2/\sigma^2) \leq e$; for $\zeta \sim \mathcal{N}(0, s_g^2 I_d)$ this holds iff $\sigma^2 \geq 2s_g^2/(1 - e^{-2/d})$, so the certified parameter is $\sigma = s_g\sqrt{2/(1 - e^{-2/d})} \approx 1.601$ at $d = 40$, and *this calibrated value*—not the generation scale—is what the batch rule of Practical ARC-SG consumes. Every method is charged one call per sampled component per iteration; the nominal budget is $2 \cdot 10^5$ calls in both families (the overparameterized family of Appendix J.1 uses $4 \cdot 10^5$). The historical driver starts a whole mini-batch whenever its counter is below the nominal budget and, for a proximal level, completes the subsequent adaptive cap measurement; hence its final counter can exceed the nominal budget. All reported $25\%/50\%/100\%$ endpoints and plotted grid values are obtained by linear interpolation of $\log_{10}$ gap at the nominal call count, rather than by substituting the post-budget point.

All curves report the geometric mean over evaluation seeds $0, \ldots, 9$. For each seed the code independently regenerates the orthogonal rotation, root, and oracle-noise stream; methods at the same seed use the same regenerated instance, so the final top-two comparison is paired by seed (although different algorithms naturally consume their random streams differently). Baseline selection uses separate validation seeds $101, 102, 103$. The band is mean $\pm$ one standard deviation of $\log_{10}$ gap across the 10 runs; gaps are floored at 10^{-16} before averaging, so "reaching the floor" means reaching this numerical limit, not an exact zero. In addition to the mean curves, the final-gap standard deviation of the $\log_{10}$ gap over the 10 seeds is reported (it lies between ≈ 0.07 and ≈ 0.23 for the main methods), endpoint gaps at 25%, 50% and 100% of the budget are recorded, and for the two best methods (PJ-SGD and Practical ARC-SG) the paired per-seed difference of the final $\log_{10}$ gaps is computed: -0.31 ± 0.23 in (a) and -0.35 ± 0.14 in (b) (PJ-SGD minus Practical ARC-SG, negative on all 10 seeds in both families). A guard clips cosh-arguments at 500 in all oracles and function evaluations; an activation counter confirmed it never fired in any reported run (it exists to keep diverging configurations, encountered only during baseline tuning, finite).

Baselines, grids, and the tuning protocol. SGD-GS: $\gamma_t = \frac{1}{2}(L_0 + L_1\,\|\hat{g}_t\|)^{-1}$, batch 4 (batch $\lceil\rho\rceil = 400$ in the overparameterized family, where $\rho = N^2$ is the almost-sure strong-growth constant of Appendix J.1). PJ-SGD: the same recursion with Polyak–Juditsky averaging over the second half of the iterates, no additional free parameters. DS-SGD: $x_{t+1} = x_t - \gamma_0 \hat{g}_t/\sqrt{t+1}$ with γ_0 tuned, batch 4 in the additive-noise families and batch $\lceil\rho\rceil = 400$ in the overparameterized family. Clip-SGD: $x_{t+1} = x_t - \gamma\,\mathrm{clip}(\hat{g}_t, c)$ with $\gamma \in \{\frac{10^{-3}}{4}, \frac{10^{-2}}{4}, \frac{10^{-1}}{4}, \frac{1}{4}\}$ and $c \in \{2, 20, 200\}$, batch 4. Acc-blind: the Nesterov-type recursion $y_t = x_t + \frac{t}{t+3}(x_t - x_{t-1})$, $x_{t+1} = y_t - \mathrm{scale} \cdot \hat{g}_t/L_{\mathrm{blind}}$ with $L_{\mathrm{blind}} = L_0 + L_1\,\|\nabla f(x^0)\|$ and $\mathrm{scale} \in \{1, 4, 16\}$ tuned, batch 4 (we deliberately do not call this AC-SA: it is a generic accelerated template, which is the point of the comparison—a generalized-smoothness-blind accelerated method would naturally use a heuristic constant of order L_{blind}, which is *not* a certified sublevel bound; see Section I). ClipSSTM: the accelerated similar-triangles template with clipped gradients, $L \in \{4, 40, 400, 4000\}$ and clipping level in $\{2, 20, 200\}$, no restarts, batch 4. M-SGD-SGC (momentum SGD under the strong growth condition; used only in Appendix J.1): heavy-ball SGD with $\gamma \in \{10^{-2}, 10^{-1}, 1\}/(4\rho)$ and $\beta \in \{0.5, 0.9, 0.95\}$ tuned, batch $\lceil\rho\rceil$. Selection uses 3 validation seeds and the anytime objective (mean $\log_{10}$ gap over calls ≥ 5000), evaluated *over the full budget of the corresponding problem*: tuning on a shorter horizon systematically selected configurations of ClipSSTM and M-SGD-SGC that looked best early and destabilized or diverged later, which would have made the comparison unfairly bad for the baselines. We note the protocol asymmetry explicitly: the baselines receive an oracle-tuned grid search, while Practical ARC-SG is run once with fixed constants; this favors the baselines.

Practical ARC-SG and its deviations from the theory. We run the two-phase architecture of Phase I and Phase II, orchestrated by the sole wrapper Algorithm 1, with moderate practical constants and no per-instance tuning; Algorithm 3 is an exact summary of the implementation and Table 5 lists its departures from the analyzed schedules. (1) In the additive-noise families Phase I uses $B_1 = 4$; in the overparameterized family it uses $B_1 = \lceil \rho \rceil = 400$. It hands off when an adaptive measurement $\hat{g}$ (not the inaccessible exact gradient) satisfies $\|\hat{g}\| \le \frac{3}{2}\bar{G}$, instead of the theoretical $3\bar{G}$ test. (2) Additive-family proximal levels use $\lambda_j = \max\{4L_1\hat{G}_j, \Lambda_j, \mu_{\text{floor}}\}$ with $\Lambda_j = L_0 2^{1-j}$, halved *every* level (the analyzed schedule starts at $32L_0$ and halves only at binding floor levels), $\mu_{\text{floor}} = \mu$ in the strongly convex family and 0 in the convex family; their ball radius is fixed at $1/(2L_1)$ rather than $2\hat{G}_j/\lambda_j$. (3) The inner restart runs $K = \min\{12, \max\{5, \lceil \log_2(H_{0,j}/\delta_j) \rceil\}\}$ epochs of $N_{\text{ep}} = \lceil 7\sqrt{1 + L_j/\lambda_j} \rceil + 3$ iterations. Its similar-triangles weight is $a_{t+1} = (t+2)/(2L_F)$, whereas analyzed STM uses $(t+1)/(4L_F)$. (4) The transfer target is $\delta_j = \max\{\lambda_j s_j^2/2, H_{0,j}/d_j\}$, with $d_j = 1024$ when $\hat{G}_j \le \frac{1}{2}$ and $\sigma > 0$, and $d_j = 64$ otherwise; $s_j = \min\{R_0/128, \iota/L_j\}$ and $\iota = \bar{G}/256$. Thus $J_{\text{prac}} = 32$ is only a scale in the s- and ι-formulas, not the number of levels. The max deliberately makes δ_j larger than the analyzed $\min\{\lambda_j s_j^2/2, \varepsilon/8\}$. (5) The exact implemented batch heuristic is $b_k = \min\{32, 1 + \lceil 4\sigma^2/(\lambda_j \bar{\Delta}_k) \rceil\}$ with calibrated $\sigma \approx 1.601$; it omits the logarithmic and $\sqrt{\lambda_j L_F}$ factors and changes the constant of (16), so it is not merely (16) with a cap. (6) The cap update is $\hat{G}_{j+1} = \max\{10^{-9}, \min[\hat{G}_j + L_j \max\{\|c_{j+1} - c_j\|, s_j\}, \frac{8}{7}\|\hat{g}_{j+1}\| + \tau_j]\}$, with adaptive measurement batches $8, 32, 128, 256$ (a factor-four increase up to 256). (7) After at least one additive-family level, $\hat{G}_j \le 2\bar{G}$ or 40% of the nominal budget triggers the interior driver: full-step (L_0, L_1)-SGD, batch $8 + \lfloor t/40 \rfloor$, with a running average that includes its starting anchor. This driver is outside the analyzed Phase I and Phase II, as well as the finisher (Procedure 3), and hence outside the theorems; ARC-SG (levels) disables it. (8) Consequently, the finisher never fires in the main runs. (9) The overparameterized branch is different: immediately after Phase I, with no proximal levels or hop stages, it runs fixed-batch-400 half-step (L_0, L_1)-SGD and returns the last iterate without averaging. (10) The code has an unreachable-in-these-runs 2000-level safety guard. It also completes a batch started below the budget and the post-level adaptive measurement; the reported endpoints use interpolation at the nominal budget as described above. Runtime for the full study is a few minutes on one CPU core (NumPy, double precision).

Additional observations. On these instances Phase I already brings the gradient near the hand-off threshold before more than one or two proximal levels fire in the full Practical ARC-SG, so the late-stage gains visible in Figure 2 are produced by the interior averaging stage; the pure-levels ablation (final gap 8.0×10^{-3} in the convex family, versus 5.2×10^{-5} for the full method) quantifies this attribution, and the PJ-SGD baseline isolates it directly: averaging alone reaches 2.6×10^{-5}, so the levels contribute speed rather than endpoint accuracy on these instances. In the diagnostic ablation runs of Figures 3–4, Phase I tames the super-critical gradient ($\|\nabla f(x^0)\| = 5967.6 = 4050\,\bar{G}$ with the certified $\bar{G} = L_0/L_1 \approx 1.4735$) at a visibly linear rate within a few hundred calls, and the subsequent ≈ 41 levels contract the cap estimate geometrically. We do not claim a parametric statistical floor in the convex family: $\nabla^2 f(x^\star)$ is singular along the quartic directions, $\sigma^2 \operatorname{tr}(\nabla^2 f(x^\star)^{-1})$ is undefined, and the interior stage is simply still descending at the budget end; in the strongly convex family the curve is likewise still decreasing, so the budget, not a floor, determines the final numbers. Finally, the centralized guard counter—incremented inside the stochastic oracle *and* inside every deterministic function/gradient evaluation used for logging—reported zero cosh-argument clips across all reported runs, so no reported trajectory was affected by the numerical safeguard.

Matched ablations. Two further curves decompose Practical ARC-SG into its components under the same 10 independently regenerated instance–noise seeds (geometric-mean gaps at 25%/50%/100% of the budget): PJ-SGD (matched), the interior PJ driver run from x^0 with the same stepsize $1/(L_0 + L_1\|\hat{g}\|)$, growing batch $8 + \lfloor t/40 \rfloor$, an average that includes the starting anchor, and the same budget; and Phase I + PJ (matched), the same driver started from the Phase I output. Convex family: PJ-SGD (matched) reaches $2.5 \times 10^{-4}/1.2 \times 10^{-4}/5.8 \times 10^{-5}$, Phase I + PJ (matched) reaches $1.8 \times 10^{-4}/9.3 \times 10^{-5}/5.5 \times 10^{-5}$, and Practical ARC-SG (reference) reaches $1.9 \times 10^{-4}/8.0 \times 10^{-5}/5.2 \times 10^{-5}$. Strongly convex family: $5.4 \times 10^{-3}/2.5 \times 10^{-3}/1.2 \times 10^{-3}$ for PJ-SGD (matched), $2.4 \times 10^{-4}/1.1 \times 10^{-4}/5.5 \times 10^{-5}$ for Phase I + PJ (matched), and

Algorithm 3 Practical ARC-SG (complete practical implementation run in Section I and Appendix J.1)

Require: x^0, R_0, certified (L_0, L_1), calibrated σ, μ_{floor} ($= \mu$ for the strongly convex additive family and 0 for the convex family), oracle budget; fixed constants, no per-instance tuning

Phase I: $x \leftarrow x^0$; $B_1 \leftarrow 4$ for additive noise and $B_1 \leftarrow \lceil \rho \rceil = 400$ for the overparameterized family
while the call counter is below the nominal budget **do**
  $\hat{g} \leftarrow \mathsf{GradEst}(x, B_1)$
  if $\|\hat{g}\| \le \frac{3}{2}\bar{G}$ **then**
    $(\hat{g}_m, \tau) \leftarrow$ adaptive measurement at x (batches $8, 32, 128, 256$)
    if $\|\hat{g}_m\| \le \frac{3}{2}\bar{G}$ **then**
      break
    end if
    $\hat{g} \leftarrow \hat{g}_m$
  end if
  $x \leftarrow x - \hat{g}/[2(L_0 + L_1\|\hat{g}\|)]$
end while
If no measurement was made before budget exhaustion, make one now; $c_1 \leftarrow x$, $\hat{G}_1 \leftarrow \max\{\frac{8}{7}\|\hat{g}_m\| + \tau, \frac{1}{4}\bar{G}\}$, $\Lambda_1 \leftarrow L_0$
if the family is overparameterized **then**
  while the call counter is below the nominal budget **do**
    $\hat{g} \leftarrow \mathsf{GradEst}(c, 400)$; $c \leftarrow c - \hat{g}/[2(L_0 + L_1\|\hat{g}\|)]$
  end while
  return c (last iterate; no levels, hops, or averaging)
end if
Additive-family Phase II: execute at least one level while the counter is below budget (driver safety guard: at most 2000 levels)
$\lambda_j \leftarrow \max\{4L_1\hat{G}_j, \Lambda_j, \mu_{\text{floor}}\}$; $L_j \leftarrow 4(L_0 + L_1\hat{G}_j)$; $\Lambda_{j+1} \leftarrow \Lambda_j/2$
$s_j \leftarrow \min\{R_0/128, \iota/L_j\}$, $\iota = \bar{G}/256$; $H_{0,j} \leftarrow \hat{G}_j^2/\lambda_j$; $\delta_j \leftarrow \max\{\lambda_j s_j^2/2, H_{0,j}/d_j\}$, where $d_j = 1024$ if $\hat{G}_j \le \frac{1}{2}, \sigma > 0$, and 64 otherwise
Run $K = \min\{12, \max\{5, \lceil \log_2(H_{0,j}/\delta_j) \rceil\}\}$ epochs, each with $N_{\text{ep}} = \lceil 7\sqrt{1 + L_j/\lambda_j} \rceil + 3$ iterations on $\mathbb{B}(c_j, 1/(2L_1))$
In epoch k, $\bar{\Delta}_k = H_{0,j}2^{-k}$ and $b_k = \min\{32, 1 + \lceil 4\sigma^2/(\lambda_j\bar{\Delta}_k) \rceil\}$; inside STM, $a_{t+1} = (t+2)/(2L_F)$ and every z-update is projected onto the single ball
Set c_{j+1} to the final epoch output; make the adaptive measurement $(\hat{g}_{j+1}, \tau_j)$
$\hat{G}_{j+1} \leftarrow \max\{10^{-9}, \min[\hat{G}_j + L_j \max\{\|c_{j+1} - c_j\|, s_j\}, \frac{8}{7}\|\hat{g}_{j+1}\| + \tau_j]\}$
Repeat levels unless $\hat{G}_{j+1} \le 2\bar{G}$ or at least 40% of the nominal budget has been spent
Additive-family interior stage (only after that trigger): $y_0 \leftarrow c_{j+1}$, $\bar{y}_0 \leftarrow y_0$
for $t = 0, 1, \ldots$ while the call counter is below the nominal budget **do**
  $b_t \leftarrow 8 + \lfloor t/40 \rfloor$; $\hat{g}_t \leftarrow \mathsf{GradEst}(y_t, b_t)$; $y_{t+1} \leftarrow y_t - \dfrac{\hat{g}_t}{L_0 + L_1\|\hat{g}_t\|}$
  $\bar{y}_{t+1} \leftarrow \big((t+1)\bar{y}_t + y_{t+1}\big)/(t+2)$
end for
return $\bar{y}_t$ (the average includes the starting anchor)

$2.3 \times 10^{-4}/1.1 \times 10^{-4}/5.3 \times 10^{-5}$ for Practical ARC-SG. The reading is two-sided: on the convex family the raw PJ driver already captures nearly all of the practical method's accuracy, and Phase I and the levels buy only small early-budget gains; on the strongly convex family the raw driver stalls at 1.2×10^{-3}, Phase I is essential (2.4×10^{-4} already at 25% of the budget), and the proximal levels add little over Phase I plus the matched driver (5.3×10^{-5} versus 5.5×10^{-5} at the budget end).

Sensitivity. Table 6 reports final geometric-mean gaps (evaluation seeds $0, 1, 2$, convex family) under perturbations of the Gaussian generation scale s_g and of the initial-distance scale about the default configuration. At $d = 40$ the corresponding certified Assumption 3 parameter is $\sigma = 6.4037749\, s_g$: the three noise settings below therefore use calibrated $\sigma \in \{0.4002359, 1.6009437, 6.4037749\}$, not

Component	Theoretical ARC-SG	Practical ARC-SG
Phase I batch	B_1 of (19)	4 (additive); 400 (overparameterized)
Phase I threshold	$3\bar{G}$	$1.5\bar{G}$
Floor schedule Λ_j	$32L_0$, halved only at binding floor levels	$L_0 \cdot 2^{1-j}$, halved every level
Ball radius	$2\hat{G}_j/\lambda_j$	$1/(2L_1)$, fixed
Epochs K	$\lceil \log_2(H_0/\delta_{\mathrm{lvl}}) \rceil_+$	$\min\{12, \max\{5, \lceil \log_2(H_{0,j}/\delta_j) \rceil\}\}$
Iterations per epoch	$N_{\mathrm{ep}} = \lceil 24\sqrt{L_F/\lambda} \rceil$	$N_{\mathrm{ep}} = \lceil 7\sqrt{\kappa_j} \rceil + 3$
STM weight	$a_{t+1} = (t+1)/(4L_F)$	$a_{t+1} = (t+2)/(2L_F)$
Transfer target δ_j	$\min\{\lambda_j s_j^2/2,\ \varepsilon/8\}$	$\max\{\lambda_j s_j^2/2,\ H_{0,j}/d_j\}$
Batches	b_k of (16)	$\min\{32, 1 + \lceil 4\sigma^2/(\lambda_j \bar{\Delta}_k) \rceil\}$
Cap update	certified bookkeeping (Lemma 12)	heuristic observed displacement/measurement, floor 10^{-9}
Additive interior stage	absent	full-step SGD, $b_t = 8 + \lfloor t/40 \rfloor$, average includes anchor
Overparameterized branch	hop stages of Algorithm 2	post-Phase-I half-step SGD, fixed batch 400, last iterate
Budget boundary	schedules stop within declared call bound	started batch and post-level measurement complete; interpolate at nominal budget
Driver level guard	schedule cap J^{sch} then certified finisher	safety guard at 2000 levels (not reached)
Finisher	Procedure 3	never fires
Projections	exact two-ball (Remark 5)	single-ball (computational-model deviation)

Table 5: Component-by-component deviations of Practical ARC-SG from the analyzed Phase I and Phase II, together with the finisher (Procedure 3). None of the practical choices in the third column is covered by Theorems 1–2; the single-ball projection is a computational-model deviation rather than a change to a quantity entering a theorem (projection calls are not part of the oracle complexity, Remark 5).

Configuration	Practical ARC-SG	PJ-SGD	Clip-SGD
$s_g = 0.25$ ($\sigma = 1.6009$; default)	3.8×10^{-5}	2.5×10^{-5}	3.9×10^{-4}
$s_g = 0.0625$ ($\sigma = 0.4002$)	1.2×10^{-5}	4.9×10^{-6}	9.6×10^{-5}
$s_g = 1.0$ ($\sigma = 6.4038$)	2.3×10^{-4}	3.0×10^{-4}	5.7×10^{-3}
distance scale 0.3	3.8×10^{-5}	2.5×10^{-5}	3.9×10^{-4}
distance scale 3.0	3.8×10^{-5}	2.5×10^{-5}	3.9×10^{-4}

Table 6: Sensitivity on the convex family: final geometric-mean gaps over 3 independently regenerated instance–noise seeds. Noise rows give both the Gaussian generation scale s_g and its calibrated Assumption 3 parameter σ.

$\{0.0625, 0.25, 1\}$. Clip-SGD ranks third throughout, but the top two methods swap at high noise: PJ-SGD is best at the default and low-noise settings, whereas Practical ARC-SG is best at $s_g = 1$. The distance-scale rows agree only at the displayed two-significant-digit precision: for scales 0.3 and 3.0, the unrounded Practical ARC-SG gaps are respectively 3.8273×10^{-5} and 3.7853×10^{-5}, and the Clip-SGD gaps are 3.8834×10^{-4} and 3.8892×10^{-4}; the underlying objective at a fixed seed is unchanged, but the initial points and trajectories are distinct.

Projection, wall-clock, and retained artifacts. Practical ARC-SG uses single-ball projections only: the two-ball intersection of Remark 5 never arises because the practical ball radius is fixed at $1/(2L_1)$. The run counters report 90 projection calls in the convex and strongly convex runs and 0 in the overparameterized run (no levels fire there); the ARC-SG (levels) ablation reports 6191. Coarse single-seed wall-clock timings (NumPy, a single CPU (central processing unit) core): Practical ARC-SG 0.2 s versus 1.3–2.3 s for the baselines in the convex and strongly convex families; in the overparameterized family Clip-SGD takes 4.1 s and ClipSSTM 5.0 s. The historical timing stdout was not retained, so these machine-dependent timings are not independently recoverable from the bundled artifacts; the supplied `timing` stage now writes `experiments/timing.json` on

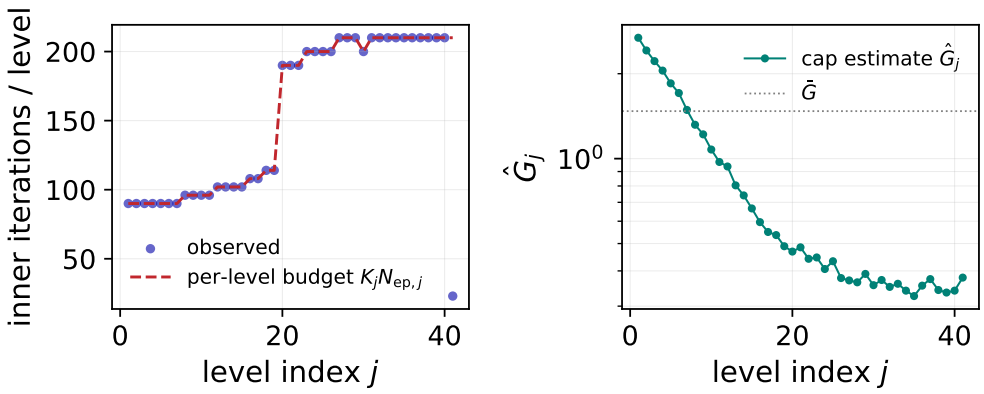


Figure 3: Diagnostics of the proximal levels (convex setting, pure accelerated levels with the interior stage disabled; the diagnostic instance starts at $\|\nabla f(x^0)\| = 4050\,\bar{G} = 5967.6$, the value logged by the run). Left: inner iterations per executed level (dots) together with the scheduled per-level budget $K_j N_{\mathrm{ep},j}$ (dashed). The implementation runs each level for exactly its scheduled budget—there is no early stopping inside a level—so the two curves agree by construction, except at the last level, truncated by the global oracle budget; the plot therefore displays the *schedule* (in particular the jump caused by the switch of the transfer-target divisor d_j from 64 to 1024 in the deep-endgame regime), not adaptive savings. Right: the cap estimate $\hat{G}_j$ contracts geometrically across levels, crossing $\bar{G}$ (dotted); the observed level count (≈ 41) is far below the schedule cap J^{sch}, which is evidence about these well-conditioned instances and not about the necessity of the worst-case $\bar{\kappa}$-factors.

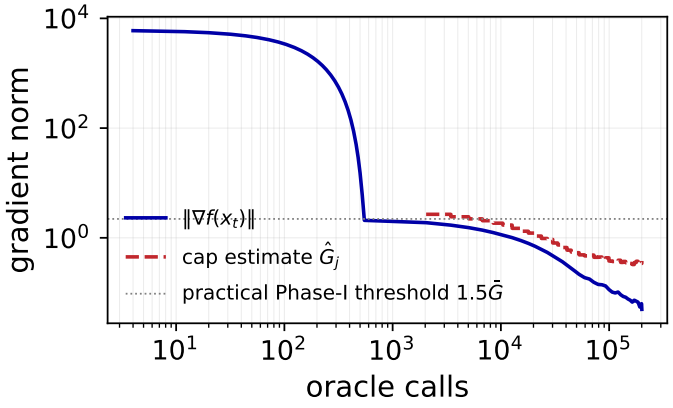


Figure 4: Gradient norm along the same pure-levels diagnostic run. The diagnostic instance scales the initial displacement along a quartic coordinate by a factor of three; since the quartic gradient is cubic in the displacement, the run starts at $\|\nabla f(x^0)\| = 5967.6 = 4050\,\bar{G}$ (logged by the run, not estimated). The two-phase structure is visible: super-critical gradients are tamed at a linear rate by Phase I (practical threshold $1.5\bar{G}$, dotted), after which proximal levels take over. The dashed curve is the *practical* cap estimate $\hat{G}_j$: on this run it lay above a high-accuracy estimate of $\|\nabla f(c_j)\|$ after every anchor and within a small constant factor of it. This is an a-posteriori observation about the implementation, not an instance of the proved invariant of Lemma 12: the practical cap update (item (6) above) replaces the certified distance-to-prox bound by the observed center displacement and is a heuristic.

rerun. Likewise, the bundled `res_*.pkl` files retain aggregate log-gap curves, endpoint statistics, and all ten final per-seed log-gaps, but not the full raw per-seed trajectories. Fresh seeded runs

reconstruct the trajectories, while the exact historical intermediate arrays and the maximum historical budget overshoot cannot be recovered from those aggregate files.

What these experiments do not test. We state this explicitly rather than leaving it to the appendix. (1) The late-stage advantage in Figure 2 is produced mainly by the interior Polyak–Juditsky averaging stage, which is *not* part of the analyzed Phase I and Phase II, or of the finisher (Procedure 3); the theory-shaped component is the levels-only ablation, which ends between SGD-GS and tuned Clip-SGD. (2) The finisher (Procedure 3) never fires on these instances, although it is exactly what makes the worst-case guarantee unconditional; targeted small instances where the finisher and the strong-growth hop stages execute are reported separately in the analyzed-schedule sanity experiment (Appendix I.1). (3) The runs use the practical schedule of Appendix I, not the analyzed constants, so the observed level counts and batch sizes carry no information about the tightness of the $\bar{\kappa}$-factors. An exact-constants run of the analyzed Phase I and Phase II, together with the finisher (Procedure 3), is computationally prohibitive: evaluated analytically from the paper's own schedule (Appendix E.1) on the convex instance at the run's accuracy, the analyzed schedule admits $J_{\mathrm{tot}} \approx 4.0 \times 10^6$ levels, and already the representative first level ($\lambda_1 \approx 68$, $L_F \approx 110$, $K = 24$ epochs, $N_{\mathrm{ep}} = 31$) has a last-epoch batch of $\approx 6.2 \times 10^9$ per iteration even with the best-case constant $C_b = 263$ of (16), for a total of $\approx 3.2 \times 10^{18}$ oracle calls ($\approx 1.6 \times 10^{13}\times$ the 2×10^5 budget; $\approx 1.2 \times 10^{20}$ with a realistic $C_b \sim 10^4$)—this is why the theory-shaped curve in Figure 2 is the levels-only ablation with moderate practical constants rather than the analyzed algorithm itself. (4) The empirical statement that the cap tracks the true gradient from above is a *posteriori* observation against a high-accuracy gradient estimate on the shown runs—the practical cap update is a heuristic, and the proved bookkeeping invariant (Lemma 12) applies to the analyzed method, not to this implementation. (5) Theorems 3–4 are not tested here; see Appendix J.1 for what the overparameterized experiment does and does not exercise. (6) Under the certified constants the overparameterized family no longer shows Practical ARC-SG competitive with tuned Clip-SGD: with the batch $\lceil \rho \rceil = 400$ that the almost-sure strong-growth condition (43) imposes, the (L_0, L_1)-stepsize methods reach only $\approx 10^{-2}$ at the budget end, while tuned Clip-SGD (batch 4) reaches the numerical floor within the first quarter of the budget; the experiment now instantiates the strong-growth assumption the theorems actually use, but it still does not exercise the hop stages of Algorithm 2.

I.1 Analyzed-schedule sanity experiment

We report a control run of the analyzed update rules—Phase I, Phase II with the full schedule of Appendix E.1, and the finisher (Procedure 3)—with the labelled best-case concentration constants $c_g = c_{\mathrm{Az}} = c_b = 1$, $c_1 = 2c_{\mathrm{Az}}$, $C_b = 263$ of (16), and $\alpha = 0.05$. The additive-Gaussian runs below simulate every batch mean exactly in distribution; the separately identified strong-growth run uses a disclosed covariance-matched approximation for very large categorical batches. A self-check verified the two-ball projection of Remark 5 against Dykstra's algorithm on 50 random cases. The test instance is the certified 1-D cosh-sum of Lemma 24(i) with scales $\{0.6, 0.8, 1.0\}$: certified pair $(L_0, L_1) = (2.1258, 1/\ln 2)$, $x^\star = 0$, $f^\star = 3$, started at $x^0 = 2.5$ with $R_0 = 2.5$; the finite-difference condition (2) was verified numerically on a grid, with maximal ratio 1.00000000 (the frontier of Lemma 24 is tight).

Deterministic oracle. With $\sigma = 0$ (a deterministic oracle, a valid degenerate case of Assumption 3) and $\varepsilon = 0.3$, the run executes 15 levels and exits via the gradient termination at the head of level 16 with true gap $f(\hat{x}) - f^\star = 4.45 \times 10^{-6} \le \varepsilon$ at 22,237 oracle calls (the exit-certified bound is 0.15). On *every* executed level the certified invariants of Lemma 12 hold: $\|\nabla f(c_j)\| \le \hat{G}_j$, $\|c_j - x^\star\| \le \bar{R}$, the certified prox lies inside the ball ($\|\hat{x}_j - c_j\| \le r_j/2$), and the R-STM output satisfies $\|c_{j+1} - \hat{x}_j\| \le \sqrt{2\delta_j/\lambda_j} \le s_j$ (each inequality is flagged individually in the log). A strongly convex companion instance (the same cosh-sum plus $0.25\,u^2$) has exact modulus $\mu = 0.6^2 + 0.8^2 + 1^2 + 0.5 = 2.5$ and certified pair $(2.6258, 1/\ln 2)$; the quadratic contribution 0.5 is merely a conservative lower bound. It executes 16 levels and exits at the head of level 17, at 23,297 calls with true gap 2.6×10^{-5}.

Noisy oracle. With $\sigma = 0.01$ the identical trajectory (same levels, same exit (a), true gap 4.7×10^{-6}) requires 2.65×10^{15} oracle calls: the batch rule (16) climbs to $\approx 1.3 \times 10^{12}$ samples per iteration on the gradient-driven levels. This quantifies, on a certified instance on which every invariant is checked

level by level, why the printed worst-case constants are not what is run at scale (cf. the practical implementation of Algorithm 3 and the deviations of Table 5).

Targeted finisher instances. The finisher (Procedure 3)—the component that never fires on the families of Figure 2—is exercised in two regimes on the same certified instance with $\sigma = 0$. (i) Under an artificial schedule cap $J^{\mathrm{sch}} = 4$ ($\varepsilon = 0.3$), control passes to the finisher, which exits correctly via exit (a) inside the finisher at its fifth level, at 5.7×10^4 oracle calls with true gap 2.4×10^{-6}. (ii) In the organic trigger attempt at $\varepsilon = 10^{-6}$, the cap falls below the drift resolution ($\hat{G}_j = 5.66 \times 10^{-6} < 32\,\iota_{\mathrm{abs}}$) and one literal finisher level is executed. The driver then aborts at its one-level iteration guard. This is *not* an exit of the paper's algorithm and carries no exit certificate; it only records that the true gap at the aborted point is 0 to machine precision.

Strong-growth algorithm. On a 3-component orthogonal interpolating cosh family (scales $\{0.08, 0.13, 0.2\}$, shared root $x^\star = 0$, $\mu = \min_k C_k^2 = 0.0064$, $\rho = N^2 = 9$, certified individual pair $(0.1298, 0.2885)$, valley half-width $\approx 70 \gg 1/L_1$), the hop-and-solve update rules and literal batch *counts* of Algorithm 2 (Stages 0–2) fire the hot stage for 4 levels and the cold stage for 1,218 hops, exiting with certified $\Delta \le \mu/(32L_1^2)$ and true gap 2.8×10^{-11}; the hop certificate $f(c_j) - f^\star \le \Delta_j$ of Lemma 21(a) held at every hop (worst ratio 0.008). To make these enormous batches executable, batches up to $5 \cdot 10^6$ use exact categorical index draws, whereas 9,521 larger batch means are drawn from a Gaussian with the same mean and exact covariance divided by the literal batch size. The call counter is still charged the literal number of samples. Thus this is a covariance-matched Gaussian approximation and a computational certificate sanity check, not an exact finite-sample reproduction of the categorical oracle.

Per-level tables of all runs are in the supplementary log file `experiments/exact_sanity.log`. The experiment code and logs accompany the submission.

J Additional experiments and limitations

J.1 Experiments in the overparameterized regime

This subsection reports the experiment corresponding to Appendix H; it is placed here, with the theory it illustrates, rather than in Section I.

Instance and constants. $d = 20$, one cosh component per rotated coordinate with scales C_k geometrically spaced in $[0.35, 1]$ ($N = 20$ components), a shared root $x^{\mathrm{int}} = Qr$ so that the family interpolates in the sense of Definition 1, $f^\star = 20$ exactly, and local strong convexity $\mu_{\mathrm{loc}} = \min_k C_k^2 = 0.1225$ exactly (the Hessian is diagonal in the rotated basis). The oracle draws one index uniformly and returns $N\nabla f_i(x)$, so the noise is multiplicative and vanishes at x^{int}. Each scaled sample Nf_i has the *Hessian-form* pair (NC_i^2, C_i), but Assumption 4 requires the finite-difference condition (2) for each individual loss, and the Hessian-form pair fails it: for $C_i = 1$, at $x = 0$, $y = 1$ (within the admissible range $|y - x| \le 1$), $|g_i'(1) - g_i'(0)| = 20\sinh(1) = 23.504 > 20$. By the frontier of Lemma 24(i) applied to $N\cosh(C_i u)$ ($B = C_i$), the valid pairs for the i-th individual loss are $L_1 \ge C_i/\ln 2$, $L_0 \ge L_1 N C_i \sinh(C_i/L_1)$, both right-hand sides being maximized at $C_i = 1$, so the tight *uniform individual* constants are

$$(L_0^{\mathrm{ind}}, L_1^{\mathrm{ind}}) = \Big(\frac{N\sinh(\ln 2)}{\ln 2},\ \frac{1}{\ln 2}\Big) = \big(\tfrac{15}{\ln 2},\ \tfrac{1}{\ln 2}\big) = (21.6404,\ 1.4427), \tag{75}$$

using $\sinh(\ln 2) = \frac{3}{4}$; necessity is attained at $C_i = 1$ (at $x = 0$, $y = \ln 2$ for L_0, and by the $x \to +\infty$ asymptotics for L_1). These—not the Hessian-form constants $(1, 1)$ of the average f—are what every (L_0, L_1)-aware method receives, exactly as Assumption 4 requires. The strong-growth constants are exact, not estimated. Because the component supports are orthogonal, $\|\nabla f(x)\|^2 = \sum_i \|\nabla f_i(x)\|^2$, hence $\mathbb{E}_i \|N\nabla f_i(x)\|^2 = \frac{1}{N}\sum_i N^2 \|\nabla f_i(x)\|^2 = N\|\nabla f(x)\|^2$: the smallest constant in the expected variant (42) is exactly $\rho = N = 20$ (an identity, so no pilot-run measurement of it is involved). The theorems, however, use the almost-sure variant (43), whose smallest uniform constant is exactly $\rho = N^2 = 400$: at single-active-coordinate points $x = x^{\mathrm{int}} + t\,q_i$ one has $\|N\nabla f_i(x)\| = N\|\nabla f(x)\|$, forcing $\rho \ge N^2$, and $\|N\nabla f_i(x)\| = N\|\nabla f_i(x)\| \le N\|\nabla f(x)\|$ for

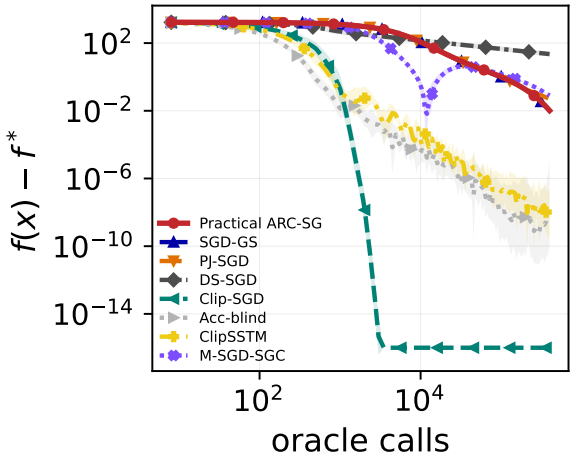


Figure 5: Overparameterized regime (interpolating cosh-family, single-sample oracle, uniform individual constants (75); Practical ARC-SG, SGD-GS, PJ-SGD, DS-SGD, and M-SGD-SGC use the strong-growth batch $\lceil \rho \rceil = 400$). Geometric mean over 10 independently regenerated instance seeds; gaps floored at 10^{-16}. No method exhibits a noise floor, as Theorem 3 predicts qualitatively: tuned Clip-SGD (batch 4) reaches the numerical floor before 25% of the budget, and Acc-blind and ClipSSTM also converge linearly, only far more slowly (3.5×10^{-8} and 1.0×10^{-8} at the budget end); the untuned (L_0, L_1)-stepsize methods Practical ARC-SG and SGD-GS (overlapping curves) descend linearly to $\approx 1.0 \times 10^{-2}$ (from $\approx 1.8 \times 10^{-1}$ at half the budget), M-SGD-SGC reaches 7.9×10^{-2}, PJ-SGD 5.0×10^{-2}, and DS-SGD ends at 2.2×10^{1}.

all x gives the matching upper bound. Everywhere the theory calls for a batch of size $\widetilde{O}(\rho)$ the runs therefore use $\lceil \rho \rceil = 400$. The budget is $4 \cdot 10^5$ oracle calls; baselines are tuned on 3 validation seeds over that full budget, and M-SGD-SGC (Vaswani et al., 2019)—momentum SGD tuned for strong growth, with batch $\lceil \rho \rceil = 400$—is added to the comparison.

What the run shows, and what it does not. With the certified individual constants (75) and the almost-sure strong-growth batch $\lceil \rho \rceil = 400$, the untuned (L_0, L_1)-stepsize methods do not reach high accuracy on this family within the budget: Practical ARC-SG and SGD-GS stand at $\approx 1.0 \times 10^{-2}$ at the budget end ($\approx 1.8 \times 10^{-1}$ at half the budget, so the linear descent is still in progress), tuned M-SGD-SGC at 7.9×10^{-2}, and PJ-SGD at 5.0×10^{-2}; tuned Clip-SGD (batch 4) reaches the 10^{-16} floor within the first quarter of the budget, while Acc-blind (3.5×10^{-8}) and ClipSSTM (1.0×10^{-8}) converge linearly but far more slowly, and DS-SGD ends at 2.2×10^{1}. Three honest qualifications are needed.

First, *the hop stages of Algorithm 2 never fire on these instances*: Phase I already lands inside the locally strongly convex basin, after which Practical ARC-SG proceeds directly to its interior driver—plain (L_0, L_1)-SGD with batch $\lceil \rho \rceil = 400$ and $\sigma = 0$—which is linearly convergent under strong growth. Consequently Figure 5 evaluates that driver, which coincides with SGD-GS up to the Phase I threshold logic (hence the two curves overlap), and *not* the hop-and-solve method of Theorems 3–4. Small targeted instances on which Stages 1–2 are genuinely exercised are reported in the analyzed-schedule sanity experiment (Appendix I.1).

Second, tuned Clip-SGD reaches the numerical floor within the first quarter of the budget, while the untuned (L_0, L_1)-stepsize methods reach only $\approx 10^{-2}$ at the budget end. This is not evidence against the theory but a direct consequence of (75): the uniform individual constants that the strong-growth theory must use are N-fold pessimistic relative to the smoothness of the average on this instance, and tuning buys back exactly that factor—and the tuned baseline (batch 4) additionally avoids the almost-sure batch $\lceil \rho \rceil = 400$ that the theory-shaped runs pay. The honest reading is that (L_0, L_1)-aware stepsizes purchase tuning-freeness and worst-case validity, not instance-optimal constants.

Third, the qualitative prediction that *is* tested—the disappearance of the noise floor under multiplicative noise—is confirmed for every method that uses a stepsize compatible with the individual constants, and fails for none; it is now illustrated under the certified individual constants and the almost-sure strong-growth batch that Theorems 3–4 actually use.

J.2 Limitations (full list)

We collect the limitations of this work in one place; each item cross-references the statement where the issue is discussed rather than re-arguing it.

(i) $\bar{\kappa}$-factors on the leading terms. The leading terms of Theorems 1–2 carry the multiplicative factors $\bar{\kappa}^{1/2}$ and $\bar{\kappa}^3$ with $\bar{\kappa} = 1 + L_1\bar{R}$; these are not proved optimal and are probably not optimal (Section G.1). Whether they can be removed, and what the minimal additive entry price is, is Open Problem 1.

(ii) External versus default initial-gap certificate. The printed $\widetilde{O}$-forms treat Δ_{cert} as an externally supplied certificate. Under the default gap certificate of Lemma 7 the Phase I entry terms scale as $(1+q)^2$ in the deterministic part and as $(1+q)^4\sigma^2 L_1^2/L_0^2$ in the statistical burn-in, where $q = L_1 R_0$ (Corollary 2; Remarks 1–2); whether the quadratic large-q dependence is sharp is open.

(iii) Exact-projection model. All theorems assume an *exact* Euclidean projection oracle onto balls and, in R-STM, onto intersections of two balls; projection calls are counted separately from stochastic-gradient oracle calls, and no projection-error (inexact-projection) analysis is provided (Remark 5, and the projection-model paragraph after Proposition 1).

(iv) Parameter knowledge. The analyzed method requires knowledge of (L_0, L_1, σ, R_0), of μ in the strongly convex case, of ρ in the strong-growth regime, and of a gap certificate Δ_{cert} for the printed $\widetilde{O}$-forms (the default certificate of Lemma 7 is available at the price of item (ii)); the full schedule is in Appendix E.1.

(v) Exact versus practical algorithms. The analyzed Phase I and Phase II, together with the finisher (Procedure 3), are *not* what the experiments run: an exact-constants run is computationally prohibitive because of the worst-case schedule constants (on the convex experimental instance, $J_{\mathrm{tot}} \approx 4 \times 10^6$ levels and $\gtrsim 3 \times 10^{18}$ oracle calls at the run's accuracy; Appendix I). The practical implementation is a distinct, clearly labelled heuristic (Appendix I) whose late-stage gain comes from a Polyak–Juditsky averaging stage that lies outside Theorems 1–2 (Section I).

(vi) Components not exercised by the main experiments. The finisher (Procedure 3) and the hot/cold hop stages of Algorithm 2 never fire on the families of Section I and Appendix J.1. In the targeted sanity runs (Appendix I.1), an artificially capped schedule enters the finisher and terminates correctly; a separate organic-trigger attempt enters it but is aborted by a driver guard after one level and is not certified. Both hop stages fire in the strong-growth check, whose 9,521 very large categorical batch means use the disclosed covariance-matched Gaussian approximation.

(vii) Almost-sure strong growth. The strong-growth theory assumes the almost-sure variant (43) of Assumption 4, which is strong; the obstruction to an in-expectation extension is discussed in Remark 13.

(viii) Theorem 4 is polynomial in $1/\varepsilon$. The convex strong-growth rate $\widetilde{O}\big(\rho\bar{\kappa} + \rho(1 + L_0\bar{R}^2/\varepsilon)^3\big)$ has no linear (i.e. $\log(1/\varepsilon)$) dependence, in contrast to the strongly convex Theorem 3. The intermediate $\bar{\kappa}^3 A + \bar{\kappa}^2 A^2$ form of the previous version is superseded by the global regularized certificate (Lemma 23); the price is the $\frac{33}{8}$ constant in L_0' and the entry-only $\bar{\kappa}$-factor.

(ix) No matching stochastic lower bounds. The ε-exponents are optimal on the smooth subclass, but we are not aware of any stochastic lower bound certifying the necessary q-dependence of the leading constants for $L_1 > 0$; all q-dependence beyond the smooth-case bounds should be read as an upper-bound artifact (Section G.1, Open Problem 1).

(x) Synthetic experiments. The experiments use synthetic $\cosh$/quartic families with constants certified for Assumption 2; they are illustrative of the *mechanism* of the method, not of its worst-case regimes or constants (Section I).