ACCELERATED STOCHASTIC METHOD UNDER (H_0, H_1)-SMOOTHNESS AND HEAVY-TAILED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop an accelerated stochastic method for convex (H_0, H_1)-smooth optimization under heavy-tailed noise. The unbiased oracle has a finite p-th noise moment, $1 < p \leq 2$, with constant, gradient-dependent, and gap-dependent terms. Our method combines accelerated updates with clipping, projection, and phase restarts. Each iteration uses a single stochastic-gradient sample, without minibatching. We prove high-probability convergence to any desired accuracy despite clipping bias. The deterministic terms retain square-root dependence on both H_0 and H_1. In our three-component noise model, strong convexity leaves a polynomial accuracy cost only for the constant component; without it, the dependence is logarithmic even for $p < 2$. More generally, for noise moments scaling as $(f(x) - f^\star)^\alpha$, the stochastic cost becomes logarithmic at $\alpha = p$ for convex objectives and $\alpha = p/2$ under strong convexity.

1 INTRODUCTION

Training modern AI models is not only stochastic; it can also be intermittently unstable. Gradient clipping was introduced as an effective response to exploding gradients in recurrent networks (Pascanu et al., 2013), and it is now a standard safeguard in large-scale learning. Yet instability remains costly at scale: the largest PaLM model experienced roughly twenty irregular loss spikes despite gradient clipping, requiring checkpoint recovery and skipped data batches (Chowdhery et al., 2023). Clipping is also not theoretically innocuous: applied to stochastic gradients, it generally creates bias whose effect depends on the threshold and on the noise model (Koloskova et al., 2023). These observations expose a tension between acceleration and robustness.

Two forms of nonuniformity make this tension particularly relevant to machine learning. First, the local curvature of a neural-network loss can vary substantially along training and correlate with the gradient scale, motivating (L_0, L_1) and more general smoothness models (Zhang et al., 2020; Li et al., 2023). Recent work on (H_0, H_1)-smoothness treats curvature as loss-dependent and shows that the admissible local step size can increase as the objective gap decreases. This model provides theoretical and empirical explanations of learning-rate warmup on language and vision models (Liu et al., 2025; Alimisis et al., 2025). Second, studies across architectures, datasets, and discriminative and generative models find strongly non-Gaussian, heavy-tailed stochastic-gradient noise, making large errors especially harmful to momentum (Simsekli et al., 2019; Battash et al., 2024). A global smoothness constant together with a state-independent noise bound can therefore be simultaneously pessimistic about both the geometry and the statistics encountered along the trajectory.

Beyond being heavy-tailed, stochastic-gradient noise can depend on the current iterate. In interpolation regimes, component gradients vanish at a common solution, while expected-smoothness conditions make their variability decrease with the objective gap (Ma et al., 2018; Gower et al., 2019). This motivates separating residual, gradient-dependent, and gap-dependent noise in our oracle model.

Existing results address important parts of this picture separately. Accelerated clipping methods handle uniform heavy-tailed noise under classical smoothness (Gorbunov et al., 2020; Nguyen et al., 2023); generalized-smoothness analyses allow bounded-variance or affine-variance noise (Li et al., 2023; Yu et al., 2025); recent high-probability acceleration under generalized smoothness treats additive norm-sub-Gaussian noise (Dvinskikh et al., 2026); and deterministic acceleration is known for the convex (H_0, H_1) class (Lobanov, 2026). The closest interpolation results under heavy tails

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.

analyze clipped or normalized SGD without accelerated high-probability guarantees for the mixed model (Lobanov & Koloskova, 2026). This leads to the central question:

Can clipping retain an accelerated optimization term while reaching arbitrary accuracy under (H_0, H_1)-smoothness and mixed heavy-tailed noise?

Our answer is affirmative for differentiable convex objectives. We first describe the geometry. With $F(x) = f(x) - f^\star$, the Hessian condition $\|\nabla^2 f(x)\| \le H_0 + H_1 F(x)$ allows high curvature at large gaps and lower curvature near the optimum (Vaswani & Babanezhad Harikandeh, 2025; Liu et al., 2025; Alimisis et al., 2025; Lobanov & Koloskova, 2026; Vaswani et al., 2026). We use the weaker first-order bound $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + (H_0 + H_1 F(x))\|y - x\|^2$ for $\|y - x\| \le r_H$, requiring only differentiability. For twice continuously differentiable convex functions, the conditions agree up to absolute constants (Lobanov, 2026). Since this bound applies only within r_H, momentum can leave its valid range; preventing this is the first difficulty.

We next describe the stochastic model. Expected smoothness bounds $\mathbb{E}_\xi\|g(x,\xi) - g(x^\star,\xi)\|^2$ by a multiple of $F(x)$ (Gower et al., 2019); related models add constant, gap-dependent, or gradient-dependent terms (Khaled & Richtarik, 2023; Faw et al., 2022; Hong & Lin, 2024), while finite-p variants allow heavy tails (Zhang & Lin, 2026). We combine them: for a fresh sample ξ and $1 < p \le 2$, $g(x,\xi)$ satisfies $\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x)$ and $\mathbb{E}_\xi\|g(x,\xi) - \nabla f(x)\|^p \le \sigma_0^p + \sigma_1^p\|\nabla f(x)\|^p + \sigma_2^p F(x)^{p/2}$. The terms describe noise that remains at the optimum, scales with the gradient, or decreases with the gap; $p < 2$ permits infinite variance. The noise scale therefore changes between momentum queries. Clipping limits large samples but creates bias; controlling both is the second difficulty.

We address both difficulties with Restarted Clipped Stochastic Accelerated Gradient (RCSAG), illustrated in Figure 1; Section 4 gives the full algorithm. Each phase starts with the gap bound $F(w) \le \Delta$ (panel (a)). Tracking the gap and distance to $x^\star$, we show that $u_t \in Q$, $F(u_t) \le 4\Delta$, and all local-model displacements are at most r_H; this fixes phasewise bounds on curvature and noise. Panel (b) forms u_t from y_{t-1} and z_{t-1} and clips its stochastic gradient; panel (c) projects $\widetilde{z}_t$ onto Q; and panel (d) averages y_{t-1} with z_t to form y_t. These bounds control random gradient errors and clipping bias without losing acceleration. With high probability, each phase halves the gap, and restarting from y_T reaches the target accuracy.

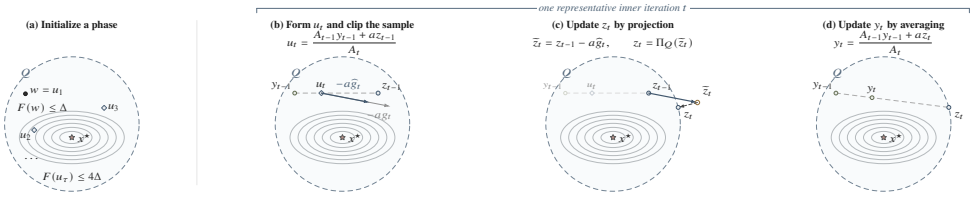


Figure 1: RCSAG: initialize, form u_t, clip, project, average, and restart after T steps.

The geometric part of our bounds retains the square-root dependence of deterministic acceleration on both curvature scales. Despite clipping bias, the method reaches any accuracy with high probability and separates the three noise components. Under strong convexity, only σ_0 forces polynomial dependence on $1/\varepsilon$; if $\sigma_0 = 0$, it is logarithmic even for $p < 2$. For noise with pth moment bounded by $F(x)^\alpha$, the cumulative cost is polynomial, logarithmic, or summable, with thresholds $\alpha = p$ for convex and $\alpha = p/2$ for strongly convex objectives. On classical smooth subclasses, the dependence on H_0, σ_0, and σ_2 matches lower bounds up to logarithms, while the σ_1 rate remains conservative.

Contributions. Our main contributions are:

- **Accelerated geometry under mixed heavy-tailed noise** (Section 5). We establish high-probability guarantees for convex and strongly convex objectives under (H_0, H_1)-smoothness and mixed heavy-tailed noise. The noise-independent geometric terms retain the square-root dependence on both H_0 and H_1 of deterministic accelerated methods.
- **Single-sample clipping without an error floor** (Sections 4 and 5). Although clipping biases each gradient estimate, RCSAG reaches any prescribed accuracy with high probability using one fresh stochastic gradient per inner iteration and no minibatching. The bounds separate noise that remains at the optimum, scales with the gradient, or decreases with the objective gap. Under strong convexity, when $\sigma_0 = 0$, the target-accuracy dependence remains logarithmic even for $p < 2$.

- **A noise-decay phase transition** (Section 5.3). For noise satisfying $\mathbb{E}_\xi\|g(x,\xi)-\nabla f(x)\|^p \lesssim F(x)^\alpha$, we characterize when its cumulative cost is polynomial, logarithmic, or summable. The critical exponent is $\alpha = p$ for convex objectives and $\alpha = p/2$ under strong convexity.

The paper focuses on convex objectives in order to expose the interaction between acceleration, nonuniform curvature, and state-dependent heavy tails with sharp objective-gap guarantees. Extending this mechanism to general nonconvex neural-network training is an important open direction.

2 Related Work

Optimization beyond a global smoothness constant. Classical acceleration uses one global smoothness constant, whereas generalized-smoothness models allow curvature to change during optimization. Under gradient-dependent (L_0, L_1)-smoothness, existing analyses explain clipping and normalization (Zhang et al., 2020), identify distinct convergence regimes for gradient descent and its variants (Lobanov et al., 2024), and extend these results to stochastic and biased gradients, as well as zeroth-order methods (Lobanov & Gasnikov, 2025). Acceleration has also been studied under broader generalized smoothness with bounded variance (Li et al., 2023), with variance depending on the current gap or gradient norm (Yu et al., 2025), and under (L_0, L_1)-smoothness with additive norm-sub-Gaussian noise (Dvinskikh et al., 2026). For the (H_0, H_1) model, Lobanov & Koloskova (2026) give non-accelerated guarantees in deterministic and interpolation settings, while Lobanov (2026) establish deterministic acceleration. Neither result gives accelerated guarantees for the mixed heavy-tailed model considered here. *Our bounds retain the accelerated square-root dependence on both H_0 and H_1 while allowing mixed finite-p heavy-tailed noise.*

Noise that decreases during optimization. A fixed global noise bound misses the fact that stochastic gradients may become more accurate near a solution. Some analyses assume that stochastic gradients approach their values at the optimum as the objective gap shrinks (Gower et al., 2019; Khaled & Richtarik, 2023), while related models combine a constant noise level with terms depending on the gradient or function value (Faw et al., 2022; Hong & Lin, 2024). Finite-p extensions allow such dependence under heavy tails (Zhang & Lin, 2026). Under interpolation, Lobanov & Koloskova (2026) show that clipped and normalized SGD can avoid a clipping-induced error floor under generalized smoothness, while Gupta et al. (2024) retain acceleration for the special case of multiplicative finite-variance noise. Thus, decreasing noise can improve convergence, but previous guarantees treat only parts of the mixed model considered here. *Our analysis handles $\mathbb{E}_\xi\|g(x,\xi)-\nabla f(x)\|^p \le \sigma_0^p + \sigma_1^p\|\nabla f(x)\|^p + \sigma_2^p F(x)^{p/2}$ without an error floor. When the noise decreases sufficiently fast near the optimum, it adds at most a logarithmic dependence on $1/\varepsilon$.*

Table 1: Convex first-order guarantees and oracle complexity.

Work	Smoothness	Tail assumption	Oracle complexity	Varying noise	Accel.	HP
Gorbunov et al. (2020)	L	finite 2nd moment	$\widetilde{O}_\varepsilon(\varepsilon^{-1/2}+\varepsilon^{-2})$	✗	✓	✓
Nguyen et al. (2023)	L	finite pth moment	$\widetilde{O}_\varepsilon(\varepsilon^{-1/2}+\varepsilon^{-q})$	✗	✓	✓
Yu et al. (2025)	(L_0, L_1)	bounded / sub-Gaussian	$\widetilde{O}_\varepsilon(\varepsilon^{-1}+\varepsilon^{-2})$	✓	✗	✓
Dvinskikh et al. (2026)	(L_0, L_1)	sub-Gaussian	$\widetilde{O}_\varepsilon(\varepsilon^{-1/2}+\varepsilon^{-2})$	✗	✓	✓
Lobanov & Koloskova (2026) (det.)	(H_0, H_1)	–	$\widetilde{O}_\varepsilon(\varepsilon^{-1})$	–	✗	–
Lobanov (2026) (det.)	(H_0, H_1)	–	$\widetilde{O}_\varepsilon(\varepsilon^{-1/2})$	–	✓	–
This work	(H_0, H_1)	finite pth moment	$\widetilde{O}_\varepsilon(\varepsilon^{-1/2}+\varepsilon^{-q})^\dagger$	✓	✓	✓

Table 1 compares the closest first-order results for convex objectives. The columns record smoothness, tail assumptions, oracle complexity, whether the noise varies along the trajectory, acceleration, and high-probability guarantees (HP). Here $q = p/(p-1)$, and $\widetilde{O}_\varepsilon$ shows only the dependence on ε, suppressing logarithms and ε-independent factors. A finite second moment still permits heavy-tailed noise, whereas sub-Gaussian noise is light-tailed. The symbols ✓, ✗, and – mean covered, not covered, and not applicable. For Yu et al. (2025), the displayed general-noise guarantee has an ε^{-1} optimization term and is therefore not marked as accelerated. In our worst-case mixed-noise entry (†), the ε^{-q} term comes only from σ_0; the σ_1 and σ_2 terms scale as $\varepsilon^{-q/2}$ in Theorem 9. *Among the listed results, ours is the only one that combines (H_0, H_1)-smoothness, finite-p tails, noise depending on the current point, acceleration, and high-probability guarantees.*

3 PROBLEM SETUP

We consider the optimization problem

$$f^\star := \min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ maps a parameter vector to a scalar objective value. Throughout the paper, $\|\cdot\|$ is the Euclidean norm. We assume that the solution set $X^\star := \arg\min_{x\in\mathbb{R}^d} f(x)$ is nonempty, fix an arbitrary $x^\star \in X^\star$ for the analysis, and write $F(x) := f(x) - f^\star$. Given an initial point x_0, an upper bound $F(x_0) \le \Delta_0$, a target accuracy $\varepsilon \in (0, \Delta_0]$, and a confidence level $\rho \in (0, 1)$, our goal is to return $\widehat{x}$ such that $F(\widehat{x}) \le \varepsilon$ with probability at least $1 - \rho$.

3.1 ASSUMPTIONS ON THE OBJECTIVE FUNCTION

We first specify the objective class considered throughout the paper.

Assumption 1 (Convexity). *The function f is differentiable and, for some $\mu \ge 0$, satisfies*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d.$$

Assumption 1 is standard in the theory of accelerated first-order methods (Nesterov, 2004). The case $\mu = 0$ gives ordinary convexity, while $\mu > 0$ gives strong convexity and is used only in Section 5.2. The latter covers, for example, convex empirical-risk objectives with an ℓ_2 regularizer.

Our main geometric assumption is the first-order form of (H_0, H_1)-smoothness used by the algorithm.

Assumption 2 ((H_0, H_1)-smoothness). *There exist $H_0, H_1 \ge 0$. Set $r_H := ((2\sqrt{3} + \sqrt{6})\sqrt{H_1})^{-1}$ when $H_1 > 0$ and $r_H := \infty$ when $H_1 = 0$. Then, for all $x, y \in \mathbb{R}^d$ satisfying $\|y - x\| \le r_H$,*

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

Assumption 2 requires only differentiability. When $H_1 = 0$, it reduces to standard L-smoothness with $L = 2H_0$. When $H_1 > 0$, the curvature scale $H_0 + H_1 F(x)$ may be large far from the optimum and approaches H_0 as the gap decreases. The upper bound is required only for $\|y - x\| \le r_H$. (H_0, H_1)-smoothness has been used to analyze line-search methods, learning-rate warmup, and first-order optimization beyond global smoothness (Vaswani & Babanezhad Harikandeh, 2025; Liu et al., 2025; Alimisis et al., 2025; Lobanov & Koloskova, 2026; Vaswani et al., 2026; Lobanov, 2026).

3.2 ASSUMPTIONS ON THE STOCHASTIC GRADIENT ORACLE

We next introduce the notation used for the stochastic oracle. At a query point x, the algorithm observes a stochastic gradient $g(x, \xi_t)$ computed from a fresh sample ξ_t; the deterministic gradient is denoted by $\nabla f(x)$. Let $\mathcal{F}_t$ be the filtration generated by the samples through iteration t, and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Each query point at iteration t is $\mathcal{F}_{t-1}$-measurable.

We impose the following condition on the conditional mean and pth moment of the stochastic gradient.

Assumption 3 (Stochastic gradient oracle). *For some $1 < p \le 2$ and $\sigma_0, \sigma_1, \sigma_2 \ge 0$, every $\mathcal{F}_{t-1}$-measurable query x admits a fresh sample $g(x, \xi_t)$ satisfying*

$$\mathbb{E}_t[g(x, \xi_t)] = \nabla f(x), \qquad \mathbb{E}_t\|g(x, \xi_t) - \nabla f(x)\|^p \le \sigma_0^p + \sigma_1^p\|\nabla f(x)\|^p + \sigma_2^p F(x)^{p/2}. \tag{2}$$

The three terms allow noise that remains at the optimum, scales with the gradient, or decreases with the objective gap. When $\sigma_1 = \sigma_2 = 0$, Assumption 3 reduces to the finite-p moment model standard in heavy-tailed stochastic optimization (Gorbunov et al., 2020; Nguyen et al., 2023); $p = 2$ gives a second-moment bound, while $p < 2$ does not require finite variance. Bounds depending on the gradient or objective gap are used to describe stochastic gradients that become more accurate during optimization (Gower et al., 2019; Faw et al., 2022; Khaled & Richtarik, 2023; Hong & Lin, 2024; Zhang & Lin, 2026). When $\sigma_0 = 0$, the bound vanishes at $x^\star$, capturing the decreasing noise found in interpolation settings (Ma et al., 2018; Lobanov & Koloskova, 2026). We write $q := p/(p-1)$. One oracle call means one evaluation of $g(x, \xi_t)$ using one fresh sample.

4 RESTARTED CLIPPED STOCHASTIC ACCELERATED GRADIENT

Classical accelerated stochastic methods are typically tuned using a global smoothness constant and a uniform noise bound. Neither scale is fixed under Assumptions 2 and 3: curvature and noise may change along the trajectory, while the upper model in Assumption 2 is guaranteed only for displacements of length at most r_H. Consequently, momentum, clipping bias, and the validity of the upper model must be controlled simultaneously. We address this difficulty through gap-halving phases. Starting from a point w satisfying $F(w) \le \Delta$, one phase is designed to return w^+ such that $F(w^+) \le \Delta/2$ with high probability. The bound Δ provides fixed upper bounds on curvature and noise during that phase; after contraction, the next phase uses the smaller bound $\Delta/2$. We implement this idea through Restarted Clipped Stochastic Accelerated Gradient (RCSAG). Algorithm 1 gives the outer restart scheme, and Procedure 1 specifies its accelerated clipped phase method.

Algorithm 1 Restarted Clipped Stochastic Accelerated Gradient (RCSAG)

Require: $x_0, \Delta_0, \varepsilon, \rho$; phase-set rule $\mathcal{Q}$; model upper bounds p, σ_i, H_i.
1: Set $w_0 \leftarrow x_0$, $S \leftarrow \lceil \log_2(\Delta_0/\varepsilon) \rceil$.
2: **for** $s = 0, \dots, S-1$ **do**
3: Set $\Delta_s \leftarrow 2^{-s}\Delta_0$, $\rho_s \leftarrow 6\rho/[\pi^2(s+1)^2]$.
4: Choose $Q_s \leftarrow \mathcal{Q}(w_s, \Delta_s)$ and set $D_s \leftarrow \operatorname{diam}(Q_s)$.
5: Update $w_{s+1} \leftarrow$ RCSAG-PHASE$(w_s, \Delta_s, Q_s, D_s, \rho_s)$.
6: **end for**
7: **return** w_S.

RCSAG (Algorithm 1) does not evaluate the objective gap during the run. It only requires an initial upper bound $F(x_0) \le \Delta_0$; all later bounds are set deterministically as $\Delta_s = 2^{-s}\Delta_0$. Appendix G shows how to construct a valid Δ_0 with high probability from stochastic gradients alone, without knowing $f^\star$, evaluating function values, or computing an exact gradient.

Algorithm 1 starts from $w_0 = x_0$ and runs $S = \lceil \log_2(\Delta_0/\varepsilon) \rceil$ phases. At phase s, w_s is the current point, $\Delta_s = 2^{-s}\Delta_0$ is an upper bound on $F(w_s)$, and ρ_s is the allowed failure probability. The rule $\mathcal{Q}$ chooses an auxiliary convex set Q_s containing w_s and $x^\star$, and D_s is its diameter. In the convex case, the same fixed set Q is used in every phase. Under strong convexity, Q_s is the shrinking ball $B(w_s, \sqrt{2\Delta_s/\mu})$. These sets control the iterates but do not constrain problem equation 1. RCSAG-Phase returns the next point w_{s+1} and, with probability at least $1-\rho_s$, halves the current gap bound: $F(w_{s+1}) \le \Delta_s/2 = \Delta_{s+1}$. Thus each successful phase preserves $F(w_s) \le \Delta_s$. Since $\sum_{s\ge 0} \rho_s \le \rho$, the final output satisfies $F(w_S) \le \varepsilon$ with probability at least $1-\rho$.

Algorithm 1 is related to the multistage AC-SA of Ghadimi & Lan (2013), which uses shrinking domains, and to R-clipped-SSTM of Gorbunov et al. (2020), which changes batch sizes and clipping thresholds across stages to exploit strong convexity. Lobanov (2026) use the same gap-halving schedule for deterministic (H_0, H_1) acceleration, while Dvinskikh et al. (2026) restart batched accelerated solves on shrinking balls under norm-sub-Gaussian noise. The outer wrapper is therefore standard. The remaining task is to realize its gap-halving step under (H_0, H_1)-smoothness and mixed heavy-tailed noise (see Assumptions 2 and 3). We now describe the phase method that does so.

4.1 ONE PHASE OF RCSAG

We now describe one call to RCSAG-Phase (Procedure 1). It starts from a point w satisfying $F(w) \le \Delta$ and uses an auxiliary convex set Q containing w and $x^\star$, with $D = \operatorname{diam}(Q) < \infty$. Its goal is to return y_T such that $F(y_T) \le \Delta/2$ with probability at least $1-\rho$.

The gap bound Δ determines phasewise bounds on curvature, gradient norm, and noise. Let $C_\star = 19 + 12\sqrt{2}$ if $H_1 > 0$ and $C_\star = 2$ otherwise; define $\mathcal{H}_\Delta := H_0 + H_1\Delta$, $G_\Delta := \sqrt{32 C_\star \Delta \mathcal{H}_\Delta}$, and $\varsigma_{p,\Delta}^p := \sigma_0^p + \sigma_1^p G_\Delta^p + \sigma_2^p (4\Delta)^{p/2}$. Here $\mathcal{H}_\Delta$ is the curvature scale for the phase, while G_Δ and $\varsigma_{p,\Delta}$ bound the gradient and noise scales along the controlled trajectory, as justified below. For $\Lambda > 0$, let $\operatorname{clip}_\Lambda(v) := v \min\{1, \Lambda/\|v\|\}$. Using D and ρ supplied by Algorithm 1, together with the bounds G_Δ and $\varsigma_{p,\Delta}$, set $\Lambda := \max\{2G_\Delta, (64\varsigma_{p,\Delta}^p D/\Delta)^{1/(p-1)}\}$, $R := D\Lambda/\Delta$, $T := \lceil 1024(1+R)\log(8/\rho) \rceil$, and $a := 5D^2/(\Delta T)$. Only T, and hence a, depends on ρ. Theorem 8 justifies these choices, and Procedure 1 below gives the resulting updates using one fresh stochastic gradient per iteration.

Procedure 1 RCSAG-PHASE

Require: w, Δ, Q, D, ρ; phase parameters Λ, T, a.
Set $y_0 = z_0 = w$, $A_0 \leftarrow D^2/\Delta$, and $A_t \leftarrow A_0 + ta$ for $t = 1, \dots, T$.
for $t = 1, \dots, T$ **do**
  Set $u_t \leftarrow (A_{t-1}y_{t-1} + az_{t-1})/A_t$.
  Query $g_t \leftarrow g(u_t, \xi_t)$ using one sample.
  Clip $\widehat{g}_t \leftarrow \mathrm{clip}_\Lambda(g_t)$.
  Update $z_t \leftarrow \Pi_Q(z_{t-1} - a\widehat{g}_t)$.
  Update $y_t \leftarrow (A_{t-1}y_{t-1} + az_t)/A_t$.
end for
return y_T.

Procedure 1 maintains a candidate output y_t and an auxiliary point $z_t \in Q$. At iteration t, it combines y_{t-1} and z_{t-1} to form the query point u_t, evaluates $g_t = g(u_t, \xi_t)$, and clips it at radius Λ to obtain $\widehat{g}_t$, so that $\|\widehat{g}_t\| \le \Lambda$. It then takes the auxiliary step $z_{t-1} - a\widehat{g}_t$, projects it onto Q to obtain z_t, and combines z_t with y_{t-1} to update y_t. The mixing fraction is a/A_t, where $A_t = A_0 + ta$; the positive A_0 keeps this fraction small during the early iterations and thereby controls the query displacements.

Procedure 1 combines two established ideas. The coupling of y_t, z_t, and u_t follows the classical accelerated construction of Nesterov (2004). Clipped-SSTM uses a related coupling with clipped minibatch gradients under L-smoothness and a uniform second-moment bound (Gorbunov et al., 2020), while Nguyen et al. (2023) analyze clipped accelerated mirror descent under a uniform finite-p noise bound. RCSAG-Phase instead adapts Λ and T to the current gap bound Δ, uses one fresh sample per iteration, and combines the positive A_0 with projection onto the auxiliary set Q to keep momentum queries within distance r_H. The individual ingredients are standard; our contribution is to combine them so that changing curvature, mixed noise, and clipping bias are controlled simultaneously without minibatching. The next subsection establishes safe queries and the bounds needed for the one-phase guarantee.

4.2 AUXILIARY RESULTS FOR ONE PHASE

The phase parameters above use G_Δ and $\varsigma_{p,\Delta}$ as bounds at every query, so we must verify them along Procedure 1. We track the potential $\Phi_t := A_t F(y_t) + \frac{1}{2}\|z_t - x^\star\|^2$, which combines the output gap with the auxiliary distance to a solution. Since $F(w) \le \Delta$, $w, x^\star \in Q$, and $A_0 = D^2/\Delta$, we have $\Phi_0 \le 3D^2/2$; also $A_T = A_0 + Ta = 6D^2/\Delta$. What remains to prove is that $\Phi_t \le 3D^2$ throughout the phase with high probability. On that event, $A_T F(y_T) \le \Phi_T \le 3D^2$, hence $F(y_T) \le \Delta/2$. The lemmas below provide the needed bounds, and the next subsection completes the argument.

Set $\mathcal{H}(x) := H_0 + H_1 F(x)$. The first lemma bounds $\|\nabla f(x)\|$ in terms of $F(x)$ and r_H.

Lemma 4 (Gradient bounds from the objective gap)**.** *Suppose Assumptions 1 and 2 hold. Let $C_\star$ be the constant defined before Procedure 1, and set $r_H^{-1} = 0$ when $H_1 = 0$. Then, for every $x \in \mathbb{R}^d$,*

$$r_H^{-1}\|\nabla f(x)\| \le C_\star \mathcal{H}(x), \qquad \|\nabla f(x)\|^2 \le 2C_\star F(x)\mathcal{H}(x). \tag{3}$$

The next lemma turns these gradient bounds into bounds at the query u_t, conditional on $\Phi_{t-1} \le 3D^2$. Appendix B shows that this condition holds throughout the phase with high probability.

Lemma 5 (Safe accelerated queries)**.** *Under Assumptions 1 and 2, run Procedure 1 with $F(w) \le \Delta$, $w, x^\star \in Q$, and the phase parameters Λ, T, a specified above. If $\Phi_{t-1} \le 3D^2$ for some $1 \le t \le T$:*

$$F(y_{t-1}) \le 3\Delta, \qquad F(u_t) \le 4\Delta, \qquad \|\nabla f(u_t)\| \le G_\Delta. \tag{4}$$

Moreover, $\|u_t - y_{t-1}\| \le r_H$ and $\|y_t - u_t\| \le r_H$.

Lemmas 4 and 5 justify G_Δ and validate both local-model steps; the key bounds are equation 3 and equation 4. Assumption 3 then gives $\mathbb{E}_t\|g(u_t, \xi_t) - \nabla f(u_t)\|^p \le \varsigma_{p,\Delta}^p$. Clipping bounds individual samples but may introduce bias, so we write $\widehat{g}_t - \nabla f(u_t) = \zeta_t + \beta_t$, where $\zeta_t := \widehat{g}_t - \mathbb{E}_t[\widehat{g}_t]$ is centered and $\beta_t := \mathbb{E}_t[\widehat{g}_t] - \nabla f(u_t)$ is the clipping bias.

Lemma 6 (Bias and noise after clipping)**.** *Suppose Assumption 3 and the conditions of Lemma 5 hold at some $t \in \{1, \dots, T\}$. Then, for $\Lambda \ge 2G_\Delta$,*

$$\|\zeta_t\| \le 2\Lambda \quad \textit{a.s.}, \qquad \|\beta_t\| \le 2\varsigma_{p,\Delta}^p \Lambda^{1-p}, \qquad \mathbb{E}_t\|\zeta_t\|^2 \le 2\varsigma_{p,\Delta}^p \Lambda^{2-p}. \tag{5}$$

The condition $\Lambda \geq 2G_\Delta$ in Lemma 6 motivates the $2G_\Delta$ term in our choice of clipping radius (given before Procedure 1). We use the bounds in equation 5 in the potential recursion below.

Lemma 7 (One-step progress). *Under the conditions of Lemma 5 at an iteration $t \in \{1, \dots, T\}$, the following bound holds for every realization of the clipped gradient $\widehat{g}_t$:*

$$\Phi_t \leq \Phi_{t-1} + a^2 \|\zeta_t + \beta_t\|^2 - a \langle \zeta_t + \beta_t, z_{t-1} - x^\star \rangle . \tag{6}$$

Together, Lemmas 4–7 (proved in Appendices A and B) provide the query, clipping, and potential bounds needed for the phase proof. In particular, equation 6 shows what the parameters defined before Procedure 1 must control: Λ limits clipping bias, while T and the resulting a control the total effect of centered noise over the phase. Appendix B shows that these choices keep $\Phi_t \leq 3D^2$ with high probability. Since $A_T = 6D^2/\Delta$, this yields $F(y_T) \leq \Delta/2$, the guarantee stated next.

4.3 One-phase guarantee

The preceding lemmas identify what must be controlled within a phase. We now state the resulting contraction guarantee and the number of stochastic-gradient samples it requires.

Theorem 8 (One-phase contraction). *Under Assumptions 1, 2, and 3, run Procedure 1 with $F(w) \leq \Delta$, $w, x^\star \in Q$, and the phase parameters Λ, T, a specified above. Then*

$$\mathbb{P}\{F(y_T) \leq \Delta/2\} \geq 1 - \rho.$$

Proof sketch. Let τ be the first $t \leq T$ with $\Phi_t > 3D^2$ (or $T+1$ if none). Up to this time, Lemmas 4 and 5 validate the local model, Assumption 3 and Lemma 6 control noise and clipping bias, and Lemma 7 gives the potential recursion. With the chosen parameters, maximal Freedman inequalities imply that, on an event of probability at least $1-\rho$, $\Phi_\tau < 3D^2$ whenever $\tau \leq T$, contradicting the definition of τ (Freedman, 1975). Hence $\mathbb{P}\{\tau \leq T\} \leq \rho$; on the complementary event, $\Phi_T \leq 3D^2$ and $A_T = 6D^2/\Delta$ gives $F(y_T) \leq \Delta/2$. The full proof is in Appendix B.

5 Main Convergence Results

Combining Theorem 8 with restarts gives an end-to-end guarantee for every prescribed $\varepsilon \in (0, \Delta_0]$ and $\rho \in (0,1)$: RCSAG reaches $F(w_S) \leq \varepsilon$ with probability at least $1-\rho$. Clipping bias imposes no fixed accuracy floor; persistent noise increases the required number of oracle calls instead. Below, $\widetilde{O}_{\rho,p}(\cdot)$ suppresses constants depending on p and logarithms in confidence and phase indices, but not polynomial dependence on the displayed problem parameters.

5.1 Convex case

Fix a closed convex set Q containing x_0 and $x^\star$, let $D := \operatorname{diam}(Q) \in (0, \infty)$, and set $S := \lceil \log_2(\Delta_0/\varepsilon) \rceil$. For $H_0, H_1 > 0$, define $J_H := \min\{S, \lceil \log_2^+(H_1\Delta_0/H_0) \rceil\}$, where $\log_2^+(x) := \max\{0, \log_2 x\}$; at the endpoints, set $J_H = 0$ if $H_1 = 0$ and $J_H = S$ if $H_0 = 0 < H_1$. Thus J_H counts the phases with $H_1\Delta_s > H_0$: before the crossover, the H_0 term $D\sqrt{H_0/\varepsilon}$ is absorbed by the H_1 contribution up to a constant; afterward, J_H is fixed and H_0 governs the later phases.

Theorem 9 (Convex). *Suppose Assumptions 1, 2, and 3 hold and $F(x_0) \leq \Delta_0$. Run Algorithm 1 with $Q_s = Q$, target $\varepsilon \in (0, \Delta_0]$, and confidence $\rho \in (0,1)$. Then, with probability at least $1-\rho$, the algorithm returns w_S satisfying $F(w_S) \leq \varepsilon$ after N_{cvx} stochastic-gradient calls with $q = p/(p-1)$:*

$$\widetilde{O}_{\rho,p}\left[S + D\left(\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H \right) + \left(\frac{\sigma_0 D}{\varepsilon} \right)^q + \left(\sigma_1 D \left[\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H^{\frac{1}{q}} \right] \right)^q + \left(\frac{\sigma_2 D}{\sqrt{\varepsilon}} \right)^q \right].$$

The bound in Theorem 9 can be read term by term. The term S comes from the constant part of the iteration budget in each restart phase. The deterministic term $D(\sqrt{H_0/\varepsilon} + \sqrt{H_1} J_H)$ preserves accelerated square-root dependence on both curvature scales (Lobanov, 2026): the H_0 part is optimal up to logarithms (Nesterov, 2004), while $J_H \leq S$ improves the previous H_1 dependence after the curvature crossover. The term $(\sigma_0 D/\varepsilon)^q$ recovers the uniform finite-p rate (Gorbunov et al., 2020;

Nguyen et al., 2023) and matches lower bounds up to logarithms (Appendix E.2). For gradient-dependent noise, Gupta et al. (2024) establish acceleration in expectation under classical smoothness for $p = 2$, whereas our high-probability bound covers mixed noise for all $1 < p \le 2$; the optimal σ_1 dependence in this broader setting remains open. Finally, $(\sigma_2 D/\sqrt{\varepsilon})^q$ captures noise that decreases with the objective gap, extending the expected-smoothness/interpolation scaling (Gower et al., 2019), and matches lower bounds up to logarithms (Appendix E.3). Hence, when $\sigma_0 = 0$, the worst stochastic accuracy dependence improves from ε^{-q} to $\varepsilon^{-q/2}$. The proof is given in Appendix C.

5.2 Strongly convex case

When $\mu > 0$ (see Assumption 1), $F(w_s) \le \Delta_s$ implies $x^\star \in Q_s = B(w_s, \sqrt{2\Delta_s/\mu})$. Thus the auxiliary ball has diameter $D_s = 2\sqrt{2\Delta_s/\mu}$ and shrinks with the gap bound. Set $S := \lceil \log_2(\Delta_0/\varepsilon) \rceil$.

Theorem 10 (Strongly convex)**.** *Suppose Assumptions 1, 2, and 3 hold with $\mu > 0$ and $F(x_0) \le \Delta_0$. Run RCSAG with the balls Q_s above, target $\varepsilon \in (0, \Delta_0]$, and confidence $\rho \in (0,1)$. Then, with probability at least $1-\rho$, it returns w_S satisfying $F(w_S) \le \varepsilon$ after N_{sc} stochastic-gradient calls:*

$$\widetilde{O}_{\rho,p}\left[\left(1+\sqrt{\frac{H_0}{\mu}}\right)S + \sqrt{\frac{H_1\Delta_0}{\mu}} + \left(\frac{\sigma_0}{\sqrt{\mu\varepsilon}}\right)^q + \left(\frac{\sigma_1}{\sqrt{\mu}}\left[\sqrt{H_0}\,S^{1/q} + \sqrt{H_1\Delta_0}\right]\right)^q + \left(\frac{\sigma_2}{\sqrt{\mu}}\right)^q S\right].$$

Recall that $q := p/(p-1)$ and, as in Theorem 9, S counts the gap-halving phases. The deterministic term $(1+\sqrt{H_0/\mu})S + \sqrt{H_1\Delta_0/\mu}$ improves the nonaccelerated bound of Lobanov & Koloskova (2026) from linear to square-root curvature dependence; its H_0 dependence is optimal up to logarithms (Nesterov, 2004). The term $(\sigma_0/\sqrt{\mu\varepsilon})^q$ attains the optimal high-probability finite-p rate (Sadiev et al., 2023) and matches our lower bound up to logarithms (Appendix E.2). The σ_1 comparison is the same as in Theorem 9, and its optimal dependence remains open. The term $(\sigma_2/\sqrt{\mu})^q S$ recovers the logarithmic $p = 2$ dependence known under expected smoothness and interpolation (Gower et al., 2019); the same dependence holds with high probability for all $1 < p \le 2$ and is sharp on a fixed-condition-number smooth subclass (Appendix E.3). Hence, when $\sigma_0 = 0$, the accuracy dependence is logarithmic even for $p < 2$. The proof is in Appendix C.

5.3 When noise decay changes the rate

To isolate how noise decay affects the accuracy dependence, in this subsection we keep conditional unbiasedness but replace the mixed moment bound in Assumption 3 with

$$\mathbb{E}_t\|g(x,\xi_t) - \nabla f(x)\|^p \le \tau_\alpha^p F(x)^\alpha, \qquad \alpha \ge 0. \tag{7}$$

The choice $\alpha = 0$ gives a uniform noise bound analogous to the σ_0 term, whereas $\alpha = p/2$ recovers the gap-dependent term $\sigma_2^p F(x)^{p/2}$ used above. Larger α means that the noise decreases faster as $F(x) \to 0$. The next result identifies when this decay changes the dependence on the desired accuracy.

Theorem 11 (Power-law noise)**.** *Suppose Assumptions 1 and 2 hold, $F(x_0) \le \Delta_0$, and the oracle is conditionally unbiased and satisfies equation 7 for some $1 < p \le 2$. Run RCSAG with target $\varepsilon \in (0, \Delta_0]$, confidence $\rho \in (0,1)$, phase noise bound $\varsigma_{p,\Delta}^p := \tau_\alpha^p (4\Delta)^\alpha$, and the corresponding choices of Q_s specified above. Then, with probability at least $1-\rho$, it returns w_S satisfying $F(w_S) \le \varepsilon$. In addition to the corresponding noise-free terms in Theorems 9 and 10, the noise contributes*

$$N_\alpha^{\mathrm{cvx}} = \widetilde{O}_{\rho,p}\big((\tau_\alpha D)^q \mathcal{K}_p(\alpha)\big), \qquad N_\alpha^{\mathrm{sc}} = \widetilde{O}_{\rho,p}\left(\left(\frac{\tau_\alpha}{\sqrt{\mu}}\right)^q \mathcal{K}_{p/2}(\alpha)\right),$$

where, for $\beta > 0$,

$$\mathcal{K}_\beta(\alpha) := \begin{cases} \varepsilon^{-(\beta-\alpha)/(p-1)}, & \alpha < \beta, \\ \log(\Delta_0/\varepsilon), & \alpha = \beta, \\ \Delta_0^{(\alpha-\beta)/(p-1)}, & \alpha > \beta. \end{cases}$$

Theorem 11 separates tail heaviness, controlled by p, from noise decay, controlled by α. In a convex phase, the stochastic cost scales as $(\tau_\alpha D)^q \Delta_s^{(\alpha-p)/(p-1)}$, so $\alpha = p$ separates polynomial, logarithmic, and accuracy-independent costs. Strong convexity gives $D_s = 2\sqrt{2\Delta_s/\mu}$ and shifts the threshold to $\alpha = p/2$; hence the power used in Assumption 3 is polynomial in the convex case but logarithmic in the strongly convex case. To our knowledge, this is the first high-probability accelerated result to identify both finite-p transitions. The proof is in Appendix D.

6 Discussion and Limitations

RCSAG requires an initial bound $F(x_0) \le \Delta_0$ but never evaluates the gap during the run; Appendix G constructs such a bound from stochastic gradients without knowing $f^\star$ or an exact gradient. In the convex case, the method also needs a bounded set Q containing x_0 and an optimizer, together with an efficient projection onto Q; standard constructions are given in Appendix G.1. Strong convexity replaces this requirement by shrinking balls determined by Δ_s and μ, which are explicit in common regularized models. The tuned Procedure 1 also uses p and upper bounds on σ_i and H_i. The horizon-driven variant in Appendix I removes them from the inner updates for a prescribed T, but the unknown parameters still determine a sufficient horizon, which the method cannot infer from observed gradients; the strongly convex wrapper also needs μ. Two theoretical questions remain open: the optimal dependence on σ_1 and matching lower bounds for the general $F(x)^\alpha$ noise model (Appendices F and D). Finally, the experiments use practical tuning and a convex head on frozen features; the finite dataset is not a literal infinite-variance oracle. Extending the theory and experiments to end-to-end nonconvex training is left for future work.

7 Experiments

Linear probing trains a prediction head while keeping the pretrained encoder fixed, and is routinely used to evaluate or adapt learned representations (Kornblith et al., 2019; Kumar et al., 2022). It gives us a convex problem built from real image features, without the additional effects of updating the encoder. We use this setup to ask a simple question: does RCSAG preserve acceleration when some examples produce unusually large gradients?

Setup and protocol. We train a kernelized linear head on frozen DINOv2 ViT-S/14 features (Oquab et al., 2024) from 128 balanced CIFAR-10 images (Krizhevsky, 2009). For the normalized regularized Gram matrix K with $\kappa(K) = 20$, we minimize $f(\alpha) = n^{-1}\sum_i \omega_i[\cosh((K\alpha)_i - y_i) - 1]$. For the realized K and weights, the problem is μ-strongly convex with $\mu = 7.65 \times 10^{-6}$ and satisfies the Hessian (H_0, H_1) bound with $H_0 = H_1 = R^2 = 2.77 \times 10^{-2}$, where $R = \max_i \|K_{i:}\|$. It also exactly interpolates: all component gradients vanish when $K\alpha^\star = y$. The fixed weights ω_i are drawn once from a Pareto distribution with shape 1.5 and create occasional large component gradients without adding oracle noise. We compare RCSAG (Algorithm 1) with Clipped SGD (Gorbunov et al., 2020), Clipped-SSTM (Sadiev et al., 2023), and a practical restarted Clipped-SSTM variant (Gorbunov et al., 2020). All methods start from zero, use the same sampled indices and one component gradient per iteration, and receive 8,192 calls. We tune 16 configurations per method on six seeds and evaluate on 30 disjoint seeds. Appendix J gives the grids, an equal-weight control, and an additional synthetic experiment.

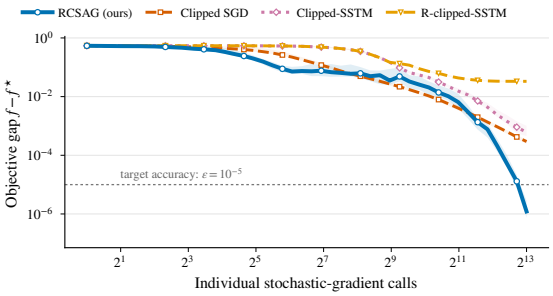


Figure 2: Median gap and 10th–90th percentile bands over 30 seeds with fixed Pareto weights.

Results. RCSAG is the only method whose final gap is below $\varepsilon = 10^{-5}$ on every held-out seed. Its median gap is 1.22×10^{-6}, compared with 2.95×10^{-4} for Clipped SGD, the next-best method, while only 0.15% of its updates are clipped. Thus rare clipping is enough to preserve fast progress on this problem; we do not claim uniform empirical superiority.

8 Conclusion

This work answers the central question affirmatively: clipping can preserve acceleration under nonuniform curvature and mixed heavy-tailed noise. RCSAG turns a gap bound into phasewise scales; projection keeps momentum queries inside the local model, while single-sample clipping and a stopped potential control errors and bias. Restarts reach any prescribed accuracy with high probability, without an error floor, while retaining square-root dependence on both curvature scales. Under strong convexity, accuracy dependence is logarithmic when $\sigma_0 = 0$, even for $p < 2$; the $F(x)^\alpha$ model exposes transitions at $\alpha = p$ and $\alpha = p/2$. Together, these results establish phasewise localization as a general design principle for accelerated optimization with evolving curvature and heavy-tailed noise without reducing either source to a uniform global worst-case bound.

AI USE STATEMENT

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

REFERENCES

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

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4780–4788, 2024.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Darina Dvinskikh, Alexander Gasnikov, Aleksandr Lobanov, and Ilgam Latypov. Localize, restart, accelerate: Stochastic optimization under generalized smoothness. *arXiv preprint arXiv:2609.06555*, 2026.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 313–355, 2022.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

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. doi: 10.1137/110848876.

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

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209, 2019.

Kanan Gupta, Jonathan W. Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yusu Hong and Junhong Lin. Revisiting convergence of adagrad with relaxed assumptions. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1727–1750, 2024.

Ahmed Khaled and Peter Richtarik. Better theory for sgd in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363, 2023.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Nikita Kornilov, Mohammad Alkousa, Eduard Gorbunov, Fedor Stonyakin, Pavel Dvurechensky, and Alexander Gasnikov. Intermediate gradient methods with relative inexactness. *Journal of Optimization Theory and Applications*, 207:62, 2025. doi: 10.1007/s10957-025-02809-y.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

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*, volume 36, 2023.

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

Aleksandr Lobanov. A few accelerated algorithms for convex optimization under (H_0, H_1)-smoothness. *arXiv preprint arXiv:2608.04884*, 2026.

Aleksandr Lobanov and Alexander Gasnikov. Power of (L_0, L_1)-smoothness in stochastic convex optimization: First- and zero-order algorithms. *arXiv preprint arXiv:2501.18198*, 2025.

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

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under (L_0, L_1)-smoothness. *arXiv preprint arXiv:2412.17050*, 2024.

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*, 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.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Ta Duy Nguyen, Thien H. Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, 2013.

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

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837, 2019.

Artem Vasin, Alexander Gasnikov, Pavel Dvurechensky, and Vladimir Spokoiny. Accelerated gradient methods with absolute and relative noise in the gradient. *Optimization Methods and Software*, 38 (6):1180–1229, 2023. doi: 10.1080/10556788.2023.2212503.

Sharan Vaswani and Reza Babanezhad Harikandeh. Armijo line-search can make (Stochastic) gradient descent provably faster. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61035–61070, 2025.

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

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

Jiayu Zhang and Tianyi Lin. Scale-invariant neural network optimization: Norm geometry and heavy-tailed noise. *arXiv preprint arXiv:2605.18528*, 2026.

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*, 2020.

APPENDIX

ACCELERATED STOCHASTIC METHOD UNDER (H_0, H_1)-SMOOTHNESS AND HEAVY-TAILED NOISE

CONTENTS

Abstract **1**

1 Introduction **1**

2 Related Work **3**

3 Problem Setup **4**

3.1 Assumptions on the objective function 4

3.2 Assumptions on the stochastic gradient oracle 4

4 Restarted Clipped Stochastic Accelerated Gradient **5**

4.1 One phase of RCSAG 5

4.2 Auxiliary Results for One Phase 6

4.3 One-phase guarantee 7

5 Main Convergence Results **7**

5.1 Convex case 7

5.2 Strongly convex case 8

5.3 When noise decay changes the rate 8

6 Discussion and Limitations **9**

7 Experiments **9**

8 Conclusion **9**

AI Use Statement **10**

References **10**

A Auxiliary Results for One Phase **15**

A.1 Proof of Lemma 4 15

A.2 Proof of Lemma 5 16

A.3 Proof of Lemma 6 17

A.4 Proof of Lemma 7 18

B Proof of Theorem 8 **19**

C Proofs of the Main Convergence Results **22**

D Power-Law Noise and Finite-Sum Derivation **25**

D.1 Power-law phase diagram 25

D.2 Finite-sum interpolation 26

E Lower Bounds on the Smooth Subclass **27**

E.1 Deterministic curvature 27

E.2 Additive finite-p noise 27

E.3 Gap/interpolation noise 29

F Multiplicative Noise: Universality versus Specialized Acceleration **31**

G Gradient-Only End-to-End Initialization **32**

G.1 Choosing the auxiliary set 32

G.2 Robust gradient estimation 32

G.3 Convex initialization 33

G.4 Strongly convex initialization 33

H Endpoints and Special Cases **34**

H.1 $p = 2$ 34

H.2 Norm-sub-Gaussian endpoint 35

H.3 Deterministic limit and other audit cases 35

I Horizon-Driven Parameter-Free Inner Method **35**

I.1 A radius-aware self-bound 35

I.2 Runtime algorithm 36

J Experimental Details and Additional Results **40**

A Auxiliary Results for One Phase

This appendix proves Lemmas 4–7. The proofs follow the order in which the results are used by Procedure 1. Lemma 4 first converts an objective-gap bound into a gradient bound. Lemma 5 then shows that the accelerated query and averaging steps remain inside the radius where Assumption 2 is valid. Lemma 6 controls the centered error and the bias created by clipping. Finally, Lemma 7 combines these ingredients in the potential recursion used to prove the one-phase contraction.

Throughout this appendix, recall that $\mathcal{H}(x) := H_0 + H_1 F(x)$, $\mathcal{H}_\Delta := H_0 + H_1\Delta$, and $C_\star = 19 + 12\sqrt{2}$ when $H_1 > 0$, whereas $C_\star = 2$ when $H_1 = 0$.

A.1 Proof of Lemma 4

Proof. Fix an arbitrary $x \in \mathbb{R}^d$ and write $F(x) = f(x) - f^\star$ and $\mathcal{H}(x) = H_0 + H_1 F(x)$. We first use one step of exactly the local-model radius to prove the estimate involving r_H^{-1}. We then use a shorter quadratic-model step to convert this estimate into the gradient self-bound. If $\nabla f(x) = 0$, both conclusions hold because their left-hand sides vanish, so below we assume $\nabla f(x) \neq 0$.

Estimate at the local-model radius. Suppose first that $H_1 > 0$. Define

$$y := x - r_H \frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

The definition gives $\|y - x\| = r_H$, so Assumption 2 applies with base point x and trial point y. Moreover, $f^\star \le f(y)$ because $f^\star$ is the minimum value. Consequently,

$$f^\star \le f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \mathcal{H}(x)\|y - x\|^2 = f(x) - r_H\|\nabla f(x)\| + \mathcal{H}(x) r_H^2. \quad (8)$$

Subtracting $f^\star$ from the last line of equation 8, rearranging, and dividing by $r_H^2 > 0$ yield

$$\frac{\|\nabla f(x)\|}{r_H} \le \frac{F(x)}{r_H^2} + \mathcal{H}(x) = \big(18 + 12\sqrt{2}\big) H_1 F(x) + \mathcal{H}(x) \le \big(19 + 12\sqrt{2}\big)\mathcal{H}(x) = C_\star \mathcal{H}(x).$$

The equality uses $r_H^{-2} = (2\sqrt{3} + \sqrt{6})^2 H_1 = (18 + 12\sqrt{2})H_1$. The last inequality follows from $H_1 F(x) \le H_0 + H_1 F(x) = \mathcal{H}(x)$. If $H_1 = 0$, then $r_H = \infty$ and the convention $r_H^{-1} = 0$ makes the first inequality in equation 3 automatic.

Gradient self-bound. We next prove the second inequality. If $\mathcal{H}(x) = 0$, then $\nabla f(x) = 0$: for $H_1 > 0$, the equality $\mathcal{H}(x) = H_0 + H_1 F(x) = 0$ implies $F(x) = 0$, so x is a minimizer of the differentiable convex function; for $H_1 = 0$, Assumption 2 has zero quadratic term globally, and the existence of a minimizer again forces the gradient to vanish. We may therefore assume that $\mathcal{H}(x) > 0$ and define

$$M_x := C_\star \mathcal{H}(x), \qquad x^+ := x - \frac{\nabla f(x)}{M_x}.$$

When $H_1 > 0$, the first inequality in equation 3 gives $\|x^+ - x\| = \|\nabla f(x)\|/M_x \le r_H$. When $H_1 = 0$, the upper model is global. Thus Assumption 2 is valid for the pair (x, x^+). Since $C_\star \ge 2$, we also have $M_x \ge 2\mathcal{H}(x)$, and hence

$$f(x^+) \le f(x) - \frac{\|\nabla f(x)\|^2}{M_x} + \mathcal{H}(x)\frac{\|\nabla f(x)\|^2}{M_x^2} \le f(x) - \frac{\|\nabla f(x)\|^2}{2M_x}. \quad (9)$$

The last inequality follows from $\mathcal{H}(x)/M_x \le 1/2$. Using $f^\star \le f(x^+)$ in equation 9 gives

$$F(x) = f(x) - f^\star \ge \frac{\|\nabla f(x)\|^2}{2M_x} = \frac{\|\nabla f(x)\|^2}{2C_\star \mathcal{H}(x)}.$$

Multiplying by $2C_\star\mathcal{H}(x)$ proves the second inequality in equation 3. □

A.2 Proof of Lemma 5

Proof. Fix $t \in \{1, \dots, T\}$ and assume $\Phi_{t-1} \le 3D^2$. We first verify that the parameter choice makes each accelerated displacement short enough for Assumption 2. We then use the potential bound to control $F(y_{t-1})$, transfer this control to the query u_t, and finally derive the gradient bound at u_t.

Weight control. Define

$$M := 32C_\star \mathcal{H}_\Delta,$$

so that $G_\Delta = \sqrt{\Delta M}$. The clipping radius satisfies $\Lambda \ge 2G_\Delta$, and therefore

$$R = \frac{D\Lambda}{\Delta} \ge 2D\sqrt{\frac{M}{\Delta}}.$$

Because $\rho \in (0,1)$ and $T = \lceil 1024(1+R)\log(8/\rho) \rceil$, we have $T \ge (5/2)R \ge 5D\sqrt{M/\Delta}$. Substituting $a = 5D^2/(\Delta T)$ and $A_0 = D^2/\Delta$ then gives

$$Ma^2 = \frac{25MD^4}{\Delta^2 T^2} \le \frac{D^2}{\Delta} = A_0 \le A_t. \tag{10}$$

Previous output gap. Since $\Phi_{t-1} = A_{t-1}F(y_{t-1}) + \frac{1}{2}\|z_{t-1} - x^\star\|^2$ and the squared-distance term is nonnegative,

$$A_{t-1}F(y_{t-1}) \le \Phi_{t-1} \le 3D^2.$$

Moreover, $A_{t-1} \ge A_0 = D^2/\Delta$. Dividing by $A_{t-1} > 0$ proves

$$F(y_{t-1}) \le 3\Delta. \tag{11}$$

First local displacement. Let $\lambda_t := a/A_t$. Equation 10 implies

$$\lambda_t^2 = \frac{a^2}{A_t^2} \le \frac{1}{MA_t} \le \frac{1}{MA_0}.$$

The set Q is convex, $y_0 = z_0 = w \in Q$, projection gives $z_j \in Q$, and the update of y_j is a convex combination of two points in Q. Hence $y_j, z_j \in Q$ for every j. Using the definition of u_t and the diameter bound for Q, we obtain

$$\begin{aligned} \|u_t - y_{t-1}\| &= \lambda_t \|z_{t-1} - y_{t-1}\| \\ &\le \lambda_t D \le \sqrt{\frac{\Delta}{M}}. \end{aligned} \tag{12}$$

If $H_1 > 0$, then $M \ge 32C_\star H_1 \Delta$, and thus

$$\frac{\Delta}{M} \le \frac{1}{32C_\star H_1} \le r_H^2.$$

The last inequality follows from $r_H^{-2} = (18 + 12\sqrt{2})H_1$ and $32C_\star \ge 18 + 12\sqrt{2}$. Therefore equation 12 gives $\|u_t - y_{t-1}\| \le r_H$. For $H_1 = 0$, the same conclusion follows from $r_H = \infty$.

Query gap and curvature. We may now apply the local upper model from y_{t-1} to u_t because equation 12 verifies its radius condition. First, equation 11 gives

$$\mathcal{H}(y_{t-1}) = H_0 + H_1 F(y_{t-1}) \le H_0 + 3H_1\Delta \le 3\mathcal{H}_\Delta.$$

Combining this inequality with the second bound in Lemma 4 yields

$$\|\nabla f(y_{t-1})\|^2 \le 2C_\star F(y_{t-1})\mathcal{H}(y_{t-1}) \le 18C_\star \Delta \mathcal{H}_\Delta.$$

Assumption 2, the Cauchy–Schwarz inequality, and equation 12 therefore give

$$\begin{aligned} F(u_t) &\le F(y_{t-1}) + \|\nabla f(y_{t-1})\|\,\|u_t - y_{t-1}\| + \mathcal{H}(y_{t-1})\|u_t - y_{t-1}\|^2 \\ &\le 3\Delta + \sqrt{18C_\star\Delta\mathcal{H}_\Delta}\sqrt{\frac{\Delta}{32C_\star\mathcal{H}_\Delta}} + 3\mathcal{H}_\Delta\frac{\Delta}{32C_\star\mathcal{H}_\Delta} \\ &= \left(3 + \frac{3}{4} + \frac{3}{32C_\star}\right)\Delta < 4\Delta. \end{aligned}$$

This proves the second gap bound in equation 4. It also implies

$$\mathcal{H}(u_t) = H_0 + H_1 F(u_t) \le H_0 + 4H_1\Delta \le 4\mathcal{H}_\Delta. \tag{13}$$

Second local displacement and query gradient. Next, subtracting the definitions of y_t and u_t gives

$$y_t - u_t = \lambda_t(z_t - z_{t-1}).$$

Both z_t and z_{t-1} belong to Q, so the same argument as in equation 12 yields $\|y_t - u_t\| \le \lambda_t D \le r_H$. Thus both displacements used later in the local-model argument are valid.

Finally, equation 13 and the definition of M give

$$\frac{\mathcal{H}(u_t)}{M} \le \frac{4\mathcal{H}_\Delta}{32C_\star\mathcal{H}_\Delta} = \frac{1}{8C_\star} \le \frac{1}{16}. \tag{14}$$

Applying the second inequality in Lemma 4 at u_t and using $F(u_t) \le 4\Delta$ together with equation 13, we obtain

$$\|\nabla f(u_t)\|^2 \le 2C_\star F(u_t)\mathcal{H}(u_t) \le 32C_\star\Delta\mathcal{H}_\Delta = G_\Delta^2.$$

Taking square roots proves $\|\nabla f(u_t)\| \le G_\Delta$ and completes the proof of equation 4. □

A.3 Proof of Lemma 6

Proof. Fix an iteration t on which the localization conclusions of Lemma 5 hold. We condition on the past, first derive the localized pth-moment bound, then verify the almost-sure clipping envelope, the clipping bias, and the conditional second moment of the centered error.

Conditioning and local moment bound. Condition on $\mathcal{F}_{t-1}$. The query u_t is then fixed because it is determined by the history before the sample ξ_t is drawn. Lemma 5 and the assumption $\Lambda \ge 2G_\Delta$ give

$$\|\nabla f(u_t)\| \le G_\Delta \le \frac{\Lambda}{2}.$$

Thus $\nabla f(u_t)$ lies in the Euclidean ball of radius Λ. Assumption 3, together with $F(u_t) \le 4\Delta$ and the displayed gradient bound, gives

$$\mathbb{E}_t\|g(u_t,\xi_t) - \nabla f(u_t)\|^p \le \sigma_0^p + \sigma_1^p G_\Delta^p + \sigma_2^p(4\Delta)^{p/2} = \varsigma_{p,\Delta}^p. \tag{15}$$

Almost-sure bound. Both $\widehat{g}_t = \mathrm{clip}_\Lambda(g(u_t,\xi_t))$ and its conditional expectation $\mathbb{E}_t[\widehat{g}_t]$ belong to the radius-Λ ball: the first statement follows from clipping, and the second follows from convexity of the ball. Therefore

$$\|\zeta_t\| = \|\widehat{g}_t - \mathbb{E}_t[\widehat{g}_t]\| \le \|\widehat{g}_t\| + \|\mathbb{E}_t[\widehat{g}_t]\| \le 2\Lambda \qquad \text{almost surely.}$$

Clipping bias. Conditional unbiasedness in Assumption 3 gives $\mathbb{E}_t[g(u_t,\xi_t)] = \nabla f(u_t)$, and hence

$$\beta_t = \mathbb{E}_t[\widehat{g}_t] - \nabla f(u_t) = \mathbb{E}_t[\widehat{g}_t - g(u_t,\xi_t)].$$

The vectors $\widehat{g}_t$ and $g(u_t,\xi_t)$ differ only on the event $\{\|g(u_t,\xi_t)\| > \Lambda\}$. On this event,

$$\|g(u_t,\xi_t) - \nabla f(u_t)\| \ge \|g(u_t,\xi_t)\| - \|\nabla f(u_t)\| > \frac{\Lambda}{2}.$$

Moreover, clipping is Euclidean projection onto the radius-Λ ball, and $\nabla f(u_t)$ belongs to that ball. The defining optimality of projection therefore gives

$$\|\widehat{g}_t - g(u_t,\xi_t)\| \le \|g(u_t,\xi_t) - \nabla f(u_t)\|.$$

Combining these facts with conditional Jensen's inequality yields

$$\begin{aligned}\|\beta_t\| &\le \mathbb{E}_t[\|g(u_t,\xi_t) - \nabla f(u_t)\|\mathbf{1}\{\|g(u_t,\xi_t) - \nabla f(u_t)\| > \Lambda/2\}] \\ &\le (\Lambda/2)^{1-p}\mathbb{E}_t\|g(u_t,\xi_t) - \nabla f(u_t)\|^p \\ &\le 2\varsigma_{p,\Delta}^p\Lambda^{1-p}.\end{aligned}$$

The second inequality uses $s\mathbf{1}\{s > c\} \le c^{1-p}s^p$ for $s, c > 0$, and the last uses equation 15 together with $2^{p-1} \le 2$ for $1 < p \le 2$.

Conditional second moment. It remains to control the conditional second moment of the centered term. Conditional expectation minimizes mean squared distance, so

$$\mathbb{E}_t\|\zeta_t\|^2 = \mathbb{E}_t\|\widehat{g}_t - \mathbb{E}_t[\widehat{g}_t]\|^2 \le \mathbb{E}_t\|\widehat{g}_t - \nabla f(u_t)\|^2.$$

Projection onto a closed convex set is nonexpansive, and $\nabla f(u_t)$ lies in the clipping ball. Hence $\|\widehat{g}_t - \nabla f(u_t)\| \le \|g(u_t, \xi_t) - \nabla f(u_t)\|$. At the same time, $\|\widehat{g}_t - \nabla f(u_t)\| \le \|\widehat{g}_t\| + \|\nabla f(u_t)\| \le 2\Lambda$. It follows that

$$\|\widehat{g}_t - \nabla f(u_t)\|^2 \le (2\Lambda)^{2-p}\|g(u_t, \xi_t) - \nabla f(u_t)\|^p.$$

Taking conditional expectations, using equation 15, and observing that $2^{2-p} \le 2$ give

$$\mathbb{E}_t\|\zeta_t\|^2 \le 2^{2-p}\varsigma_{p,\Delta}^p\Lambda^{2-p} \le 2\varsigma_{p,\Delta}^p\Lambda^{2-p}.$$

This proves all three inequalities in equation 5. □

A.4 Proof of Lemma 7

Proof. Fix an iteration t satisfying the premise of Lemma 5. We first apply the local upper model to the averaging update, then use convexity to couple the objective terms, and finally use the Euclidean projection optimality condition to control the auxiliary step.

Local-model progress. Lemma 5 gives $\|y_t - u_t\| \le r_H$, so Assumption 2 applies from u_t to y_t. Subtracting the definitions of y_t and u_t gives

$$y_t - u_t = \frac{a}{A_t}(z_t - z_{t-1}).$$

Multiplying the resulting upper-model inequality by A_t therefore yields

$$A_tF(y_t) \le A_tF(u_t) + a\,\langle \nabla f(u_t), z_t - z_{t-1}\rangle + \frac{\mathcal{H}(u_t)a^2}{A_t}\|z_t - z_{t-1}\|^2. \tag{16}$$

Equation 10 gives $Ma^2 \le A_t$, while equation 14 gives $\mathcal{H}(u_t)/M \le 1/16$. Thus the coefficient of the squared displacement in equation 16 satisfies

$$\frac{\mathcal{H}(u_t)a^2}{A_t} \le \frac{\mathcal{H}(u_t)}{M} \le \frac{1}{16}. \tag{17}$$

Convex coupling. We next relate $F(u_t)$ to the previous output gap. Convexity at u_t gives

$$F(u_t) - F(y_{t-1}) \le \langle \nabla f(u_t), u_t - y_{t-1}\rangle, \qquad F(u_t) \le \langle \nabla f(u_t), u_t - x^\star\rangle.$$

Multiplying the first inequality by A_{t-1}, the second by a, and adding them gives

$$\begin{aligned} A_tF(u_t) - A_{t-1}F(y_{t-1}) &\le \langle \nabla f(u_t), A_{t-1}(u_t - y_{t-1}) + a(u_t - x^\star)\rangle \\ &= a\,\langle \nabla f(u_t), z_{t-1} - x^\star\rangle. \end{aligned} \tag{18}$$

The equality uses $A_tu_t = A_{t-1}y_{t-1} + az_{t-1}$, which is the definition of the query point. Combining equation 16, equation 17, and equation 18 yields

$$A_tF(y_t) \le A_{t-1}F(y_{t-1}) + a\,\langle \nabla f(u_t), z_t - x^\star\rangle + \frac{1}{16}\|z_t - z_{t-1}\|^2. \tag{19}$$

Projection inequality. It remains to control the inner product in equation 19. Since $z_t = \Pi_Q(z_{t-1} - a\widehat{g}_t)$ and $x^\star \in Q$, the optimality condition for Euclidean projection gives

$$\langle z_t - z_{t-1} + a\widehat{g}_t, x^\star - z_t\rangle \ge 0.$$

Rearranging this inequality and applying the identity $2\,\langle z_{t-1} - z_t, z_t - x^\star\rangle = \|z_{t-1} - x^\star\|^2 - \|z_t - x^\star\|^2 - \|z_t - z_{t-1}\|^2$ give

$$a\,\langle \widehat{g}_t, z_t - x^\star\rangle \le \frac{1}{2}\|z_{t-1} - x^\star\|^2 - \frac{1}{2}\|z_t - x^\star\|^2 - \frac{1}{2}\|z_t - z_{t-1}\|^2. \tag{20}$$

Potential recursion. Substitute $\nabla f(u_t) = \widehat{g}_t - (\zeta_t + \beta_t)$ into equation 19, use equation 20, and add $\frac{1}{2}\|z_t - x^\star\|^2$ to both sides. Splitting $z_t - x^\star = (z_{t-1} - x^\star) + (z_t - z_{t-1})$ then gives

$$\Phi_t \le \Phi_{t-1} - a\,\langle \zeta_t + \beta_t, z_{t-1} - x^\star\rangle - a\,\langle \zeta_t + \beta_t, z_t - z_{t-1}\rangle - \left(\frac{1}{2} - \frac{1}{16}\right)\|z_t - z_{t-1}\|^2.$$

Young's inequality with coefficients 1 and $1/4$ gives

$$-a\,\langle \zeta_t + \beta_t, z_t - z_{t-1}\rangle \le a^2\|\zeta_t + \beta_t\|^2 + \frac{1}{4}\|z_t - z_{t-1}\|^2.$$

After this substitution, the remaining coefficient of $\|z_t - z_{t-1}\|^2$ is $1/16 - 1/2 + 1/4 = -3/16 \le 0$. Discarding this nonpositive term proves

$$\Phi_t \le \Phi_{t-1} + a^2\|\zeta_t + \beta_t\|^2 - a\,\langle \zeta_t + \beta_t, z_{t-1} - x^\star\rangle\,,$$

which is exactly equation 6. □

B Proof of Theorem 8

We now prove the one-phase contraction stated in Theorem 8. The proof has four steps. First, the clipping threshold is chosen so that both the clipping bias and the conditional second moment of the centered clipped gradient have the required scale. Second, we stop the potential at the first exit from the localization region; the indicator of the stopped process is predictable, so the centered error terms remain martingale differences. Third, we apply a maximal Freedman inequality to the two martingale channels produced by the linear and quadratic error terms. Finally, the resulting uniform bound rules out an exit and converts the terminal potential bound into the desired objective-gap contraction.

Recall that

$$\Phi_t = A_t F(y_t) + \frac{1}{2}\|z_t - x^\star\|^2, \qquad A_0 = \frac{D^2}{\Delta}, \qquad a = \frac{5D^2}{\Delta T}.$$

Because $F(w) \le \Delta$, $y_0 = z_0 = w$, and $w, x^\star \in Q$ with $\operatorname{diam}(Q) = D$, the initial potential satisfies

$$\Phi_0 = A_0 F(w) + \frac{1}{2}\|w - x^\star\|^2 \le D^2 + \frac{1}{2}D^2 = \frac{3}{2}D^2. \tag{21}$$

Threshold choice. On the localized trajectory, Lemmas 4 and 5 together with Assumption 3 give $\mathbb{E}_t\|g_t - \nabla f(u_t)\|^p \le \varsigma_{p,\Delta}^p$. Choose

$$\Lambda_{\Delta,p} := \max\left\{2G_\Delta, \left(\frac{64\varsigma_{p,\Delta}^p D}{\Delta}\right)^{1/(p-1)}\right\}, \qquad R_{\Delta,p} := \frac{D\Lambda_{\Delta,p}}{\Delta}. \tag{22}$$

Equation 22 combines two logically separate requirements: its first threshold localizes clipping around the true gradient, and its second threshold makes the clipping bias small relative to the target gap. For the remainder of this proof, write $\Lambda := \Lambda_{\Delta,p}$ and $R := R_{\Delta,p}$. The first term in the maximum ensures $\Lambda \ge 2G_\Delta$, which is the condition needed in Lemma 6. The second term is equivalent to

$$\Lambda^{p-1} \ge \frac{64\varsigma_{p,\Delta}^p D}{\Delta}. \tag{23}$$

Consequently, on every localized iteration, Lemma 6 yields the predictable decomposition

$$\widehat{g}_t - \nabla f(u_t) = \zeta_t + \beta_t, \qquad \mathbb{E}_t\zeta_t = 0.$$

Indeed, substituting equation 23 into the two moment bounds of Lemma 6 gives

$$\|\beta_t\| \le 2\varsigma_{p,\Delta}^p \Lambda^{1-p} \le \frac{\Delta}{32D}, \tag{24}$$

$$v_t := \mathbb{E}_t\|\zeta_t\|^2 \le 2\varsigma_{p,\Delta}^p \Lambda^{2-p} = 2\varsigma_{p,\Delta}^p \Lambda\Lambda^{1-p} \le \frac{\Delta\Lambda}{32D}. \tag{25}$$

We also need an upper bound on the dimensionless threshold R. Since $\max\{x, y\} \le x + y$ for $x, y \ge 0$, the definition of Λ and the identity $G_\Delta = \sqrt{\Delta M}$, where $M = 32C_\star\mathcal{H}_\Delta$, imply

$$R = \frac{D\Lambda}{\Delta} \le \frac{2DG_\Delta}{\Delta} + \frac{D}{\Delta}\left(\frac{64\varsigma_{p,\Delta}^p D}{\Delta}\right)^{1/(p-1)} = 2D\sqrt{\frac{M}{\Delta}} + 64^{1/(p-1)}\left(\frac{\varsigma_{p,\Delta} D}{\Delta}\right)^q, \tag{26}$$

where $q = p/(p-1)$ and hence $1 + 1/(p-1) = q$.

Choice of the horizon. Set

$$\ell_\rho := \log \frac{8}{\rho}, \qquad T := \lceil 1024(1+R)\ell_\rho \rceil . \tag{27}$$

The deterministic part of R gives $R \geq 2D\sqrt{M/\Delta}$. Since $\ell_\rho > 1$, equation 27 in particular implies $T \geq 5D\sqrt{M/\Delta}$. Therefore

$$Ma^2 = \frac{25MD^4}{\Delta^2 T^2} \leq \frac{D^2}{\Delta} = A_0 \leq A_t. \tag{28}$$

Condition equation 28 is precisely the weight condition used in Lemmas 5 and 7.

Stopped martingale channels. Set the envelope level $B_{\mathrm{env}} := 3D^2$ and define the stopping time

$$\tau := \inf\{t \in \{1, \dots, T\} : \Phi_t > B_{\mathrm{env}}\}, \qquad \inf \varnothing := T + 1. \tag{29}$$

The stopping rule in equation 29 localizes every stochastic estimate to iterations for which the safe-query lemma is already valid. For every t, let $I_t := \mathbf{1}\{t - 1 < \tau\}$. The event $\{t - 1 < \tau\}$ is determined by $\Phi_0, \dots, \Phi_{t-1}$, so I_t is $\mathcal{F}_{t-1}$-measurable. Moreover, $I_t = 1$ implies $\Phi_{t-1} \leq B_{\mathrm{env}}$, and hence Lemmas 5–7 and bounds equation 24–equation 25 are valid at iteration t.

Multiplying equation 6 by I_t and using $\|\zeta_t + \beta_t\|^2 \leq 2\|\zeta_t\|^2 + 2\|\beta_t\|^2$, we obtain

$$\begin{aligned} I_t(\Phi_t - \Phi_{t-1}) \leq{}& 2a^2 I_t v_t + 2a^2 I_t(\|\zeta_t\|^2 - v_t) + 2a^2 I_t \|\beta_t\|^2 \\ & - aI_t \langle \zeta_t, z_{t-1} - x^\star \rangle - aI_t \langle \beta_t, z_{t-1} - x^\star \rangle . \end{aligned} \tag{30}$$

The reason for adding and subtracting v_t is that the second term on the right is now centered conditionally on $\mathcal{F}_{t-1}$. The fourth term is centered for the same reason, while the remaining three terms are predictable or deterministic upper bounds.

Predictable contributions. Because $z_{t-1}, x^\star \in Q$, we have $\|z_{t-1} - x^\star\| \leq D$. Using equation 24, $Ta = 5D^2/\Delta$, and $I_t \leq 1$ gives

$$\sum_{t=1}^{T} aI_t D\|\beta_t\| \leq TaD\frac{\Delta}{32D} = \frac{5}{32}D^2, \tag{31}$$

$$2\sum_{t=1}^{T} a^2 I_t \|\beta_t\|^2 \leq 2Ta^2 \frac{\Delta^2}{32^2 D^2} = \frac{50}{1024T}D^2. \tag{32}$$

Similarly, equation 25 and $R = D\Lambda/\Delta$ yield

$$2\sum_{t=1}^{T} a^2 I_t v_t \leq 2Ta^2 \frac{\Delta\Lambda}{32D} = \frac{50}{32}\frac{R}{T}D^2. \tag{33}$$

Equations 27, 32, and 33 imply

$$\frac{50}{1024T}D^2 + \frac{50}{32}\frac{R}{T}D^2 < \frac{1}{100}D^2. \tag{34}$$

Indeed, $T \geq 1024\ell_\rho$ and $R/T \leq [1024\ell_\rho]^{-1}$, while $\ell_\rho > 1$.

Centered contributions. For $j \leq T$, define the stopped scalar martingales

$$M_j^{(1)} := -\sum_{t=1}^{j} aI_t \langle \zeta_t, z_{t-1} - x^\star \rangle , \qquad M_j^{(2)} := 2a^2 \sum_{t=1}^{j} I_t(\|\zeta_t\|^2 - v_t).$$

Both are martingales because I_t and z_{t-1} are $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}_t\zeta_t = 0$, and $v_t = \mathbb{E}_t\|\zeta_t\|^2$. For the first channel, $\|\zeta_t\| \leq 2\Lambda$ gives

$$|\Delta M_t^{(1)}| \leq 2a\Lambda D = \frac{10D^2 R}{T}, \tag{35}$$

$$V_T^{(1)} := \sum_{t=1}^{T} \mathbb{E}_t[(\Delta M_t^{(1)})^2] \leq a^2 D^2 \sum_{t=1}^{T} I_t v_t \leq \frac{25}{32}\frac{D^4 R}{T}. \tag{36}$$

For the quadratic channel, both $\|\zeta_t\|^2$ and v_t lie in $[0, 4\Lambda^2]$, so $|\|\zeta_t\|^2 - v_t| \le 4\Lambda^2$. In addition, $\|\zeta_t\|^4 \le 4\Lambda^2\|\zeta_t\|^2$. Therefore

$$|\Delta M_t^{(2)}| \le 8a^2\Lambda^2 = \frac{200D^2R^2}{T^2}, \tag{37}$$

$$V_T^{(2)} := \sum_{t=1}^{T} \mathbb{E}_t[(\Delta M_t^{(2)})^2] \le 4a^4 \sum_{t=1}^{T} I_t \mathbb{E}_t\|\zeta_t\|^4 \le 16a^4\Lambda^2 \sum_{t=1}^{T} I_t v_t \le \frac{625}{2}\frac{D^4R^3}{T^3}. \tag{38}$$

We use the following two-sided maximal form of Freedman's inequality: if a scalar martingale has increments bounded in absolute value by L and predictable quadratic variation at most V, then, for every $u > 0$,

$$\mathbb{P}\left\{\max_{j\le T}|M_j| \ge u\right\} \le 2\exp\left(-\frac{u^2}{2(V + Lu/3)}\right).$$

Take $u = D^2/8$. From equation 27,

$$\frac{R}{T} \le \frac{1}{1024\ell_\rho}, \qquad \frac{R^2}{T^2} \le \frac{1}{1024^2\ell_\rho^2}, \qquad \frac{R^3}{T^3} \le \frac{1}{1024^3\ell_\rho^3}.$$

Substitution of the linear increment and variation bounds equation 35 and equation 36, together with the quadratic bounds equation 37 and equation 38, shows, for each $i \in \{1, 2\}$,

$$\frac{u^2}{2(V_T^{(i)} + L_i u/3)} \ge \ell_\rho.$$

For example, for the linear channel this follows from

$$2\left(\frac{25}{32 \cdot 1024} + \frac{10}{24 \cdot 1024}\right) \le \frac{1}{64};$$

the quadratic channel has still smaller normalized constants because it contains 1024^{-2} and 1024^{-3}. Hence

$$\mathbb{P}\left\{\max_{j\le T}|M_j^{(i)}| > \frac{D^2}{8}\right\} \le 2e^{-\ell_\rho} = \frac{\rho}{4}, \qquad i = 1, 2. \tag{39}$$

Closing the stopping argument. Let $\mathcal{E}$ be the event on which both martingales in equation 39 remain within $D^2/8$. By a union bound, $\mathbb{P}(\mathcal{E}) \ge 1 - \rho/2 \ge 1 - \rho$. For any $j \le T$, the definition of I_t gives the telescoping identity

$$\sum_{t=1}^{j} I_t(\Phi_t - \Phi_{t-1}) = \Phi_{j\wedge\tau} - \Phi_0.$$

On $\mathcal{E}$, sum equation 30, use equation 21, equation 31, equation 34, and the two martingale bounds, and obtain, simultaneously for all $j \le T$,

$$\Phi_{j\wedge\tau} \le \frac{3}{2}D^2 + \frac{5}{32}D^2 + \frac{1}{8}D^2 + \frac{1}{8}D^2 + \frac{1}{100}D^2 < 3D^2 = B_{\text{env}}.$$

If $\tau \le T$, choosing $j = \tau$ would give $\Phi_\tau < B_{\text{env}}$, contradicting the definition of τ. Thus, on $\mathcal{E}$, necessarily $\tau = T + 1$, and therefore $\Phi_T \le B_{\text{env}}$.

Finally,

$$A_T = A_0 + Ta = \frac{D^2}{\Delta} + \frac{5D^2}{\Delta} = \frac{6D^2}{\Delta}.$$

Since the distance term in Φ_T is nonnegative,

$$F(y_T) \le \frac{\Phi_T}{A_T} \le \frac{\Delta}{2}.$$

This conclusion holds on an event of probability at least $1 - \rho$, which proves Theorem 8.

C Proofs of the Main Convergence Results

This section derives Theorems 9 and 10 from the one-phase result. We first obtain a common bound for the cost of a phase with gap level Δ. We then sum that bound with a fixed set diameter for the convex result and with shrinking diameters for the strongly convex result. The only slightly nonstandard point is the H_1 crossover: after $H_1\Delta_s$ falls below H_0, the gap-dependent curvature must no longer be charged once per remaining phase.

Cost of one phase. By equation 26, $M = 32C_\star\mathcal{H}_\Delta$, and $T = \lceil 1024(1 + R_{\Delta,p})\log(8/\rho)\rceil$, the cost of one phase is bounded in terms of $\varsigma_{p,\Delta}$. To make this dependence explicit, recall that

$$\varsigma_{p,\Delta}^p = \sigma_0^p + \sigma_1^p G_\Delta^p + \sigma_2^p(4\Delta)^{p/2}, \qquad G_\Delta^p = \left(32C_\star\Delta\mathcal{H}_\Delta\right)^{p/2}.$$

Since $q/p = 1/(p-1) \geq 1$, the elementary inequality $(u+v+w)^r \leq 3^{r-1}(u^r + v^r + w^r)$, valid for $r \geq 1$ and nonnegative u, v, w, gives

$$\left(\frac{\varsigma_{p,\Delta}D}{\Delta}\right)^q \leq 3^{q/p-1}\left[\left(\frac{\sigma_0 D}{\Delta}\right)^q + 32^{q/2}\left(\sigma_1 D\sqrt{C_\star\left(\frac{H_0}{\Delta} + H_1\right)}\right)^q + 2^q\left(\frac{\sigma_2 D}{\sqrt{\Delta}}\right)^q\right].$$

Consequently, for a constant $C_p > 0$ depending only on p,

$$\begin{aligned} T \leq C_p \log\frac{8}{\rho}\Bigg[1 + D\sqrt{C_\star\left(\frac{H_0}{\Delta} + H_1\right)} + \left(\frac{\sigma_0 D}{\Delta}\right)^q \\ + \left(\sigma_1 D\sqrt{C_\star\left(\frac{H_0}{\Delta} + H_1\right)}\right)^q + \left(\frac{\sigma_2 D}{\sqrt{\Delta}}\right)^q\Bigg]. \end{aligned} \tag{40}$$

For instance, $C_p = 1024\,64^{q/p}3^{q/p-1}32^{q/2}$ is valid; these numerical factors are proof bounds, not tuning recommendations.

Restart events and geometric sums. Let $\Delta_s = 2^{-s}\Delta_0$ and $S = \lceil \log_2(\Delta_0/\varepsilon)\rceil$. We repeatedly use the following explicit sums. If $S = 0$, then $\varepsilon = \Delta_0$, the restart loop is empty, and $w_S = w_0$ already satisfies $F(w_S) \leq \varepsilon$. Thus all sums below are empty and both convergence statements are immediate. For the remainder of the restart analysis, assume $S \geq 1$. The ceiling definition implies

$$2^{S-1} < \frac{\Delta_0}{\varepsilon} \leq 2^S.$$

For every $a > 0$, substituting $\Delta_s = 2^{-s}\Delta_0$ into the increasing geometric sum gives

$$\sum_{s=0}^{S-1}\Delta_s^{-a} = \Delta_0^{-a}\sum_{s=0}^{S-1}2^{as} = \Delta_0^{-a}\frac{2^{aS}-1}{2^a-1} \leq \frac{2^a}{2^a-1}\,\varepsilon^{-a}. \tag{41}$$

The last inequality uses $2^S < 2\Delta_0/\varepsilon$, which follows from $2^{S-1} < \Delta_0/\varepsilon$. Likewise, the decreasing geometric sum is

$$\sum_{s=0}^{S-1}\Delta_s^a = \Delta_0^a\sum_{s=0}^{S-1}2^{-as} = \Delta_0^a\frac{1-2^{-aS}}{1-2^{-a}} \leq \frac{\Delta_0^a}{1-2^{-a}}. \tag{42}$$

Finally, $\sum_{s=0}^{S-1} 1 = S$ because there are exactly S phases. A summable failure allocation, for example $\rho_s = 6\rho/[\pi^2(s+1)^2]$, gives $\sum_{s\geq 0}\rho_s = \rho$. To justify the restart induction explicitly, let

$$\mathcal{E}_s := \{F(w_j) \leq \Delta_j \text{ for } j = 0, \ldots, s\}.$$

The initial assumption gives $\mathcal{E}_0$. Conditional on $\mathcal{E}_s$, the set Q_s contains w_s and $x^\star$, so Theorem 8 gives

$$\mathbb{P}(\mathcal{E}_{s+1}^c \cap \mathcal{E}_s) \leq \rho_s.$$

A union bound over $s = 0, \ldots, S-1$ therefore yields

$$\mathbb{P}(\mathcal{E}_S) \geq 1 - \sum_{s=0}^{S-1}\rho_s \geq 1 - \rho.$$

On $\mathcal{E}_S$, $F(w_S) \leq \Delta_S \leq \varepsilon$. Moreover, $\log(8/\rho_s) = O(\log(1/\rho) + \log(s+1))$; these confidence and phase-index logarithms are precisely the factors absorbed by $\widetilde{O}_{\rho,p}$.

A crossover lemma. The technical note used the valid but coarse inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, which yields an H_1 factor on every phase. The following split is strictly sharper whenever the algorithm crosses into the H_0-dominated regime. We assign the equality case $H_1\Delta_s = H_0$ to the H_0-dominated regime and use J_H as defined in Section 5.1.

Lemma 12 (H_1 crossover sums)**.** *For every $r \ge 1/2$,*

$$\sum_{s=0}^{S-1} \left(\frac{H_0}{\Delta_s} + H_1 \right)^r \le C_r \left[\left(\frac{H_0}{\varepsilon} \right)^r + H_1^r J_H \right]. \tag{43}$$

Proof. Fix $r \ge 1/2$. If $H_1 = 0$, substituting $a = r$ into equation 41 proves the claim. If $H_0 = 0$, every summand equals H_1^r and $J_H = S$, so both sides have the same dependence on $H_1^r S$. Assume now that $H_0 > 0$ and $H_1 > 0$, and split the phase indices into $I_1 = \{s : H_1 > H_0/\Delta_s\}$ and $I_0 = \{s : H_1 \le H_0/\Delta_s\}$. Since $\Delta_s = 2^{-s}\Delta_0$, the indices in I_1 satisfy $s < \log_2(H_1\Delta_0/H_0)$, so their number is at most J_H. On I_1,

$$\sum_{s \in I_1} \left(\frac{H_0}{\Delta_s} + H_1 \right)^r \le 2^r H_1^r |I_1| \le 2^r H_1^r J_H.$$

On I_0, including the equality case, $H_1 \le H_0/\Delta_s$, and hence

$$\sum_{s \in I_0} \left(\frac{H_0}{\Delta_s} + H_1 \right)^r \le 2^r H_0^r \sum_{s \in I_0} \Delta_s^{-r} \le 2^r H_0^r \sum_{s=0}^{S-1} \Delta_s^{-r} \le \frac{2^{2r}}{2^r - 1} \left(\frac{H_0}{\varepsilon} \right)^r,$$

where the last line applies equation 41 with $a = r$. Adding the two regions proves equation 43 with, for example, $C_r = \max\{2^r, 2^{2r}/(2^r - 1)\}$. □

Proof of Theorem 9. In the convex setting, the same set Q and diameter D are used in every phase. Summing the constant term in equation 40 gives S. For the deterministic curvature term, Lemma 12 with $r = 1/2$ gives

$$\sum_{s=0}^{S-1} D \sqrt{C_\star \left(\frac{H_0}{\Delta_s} + H_1 \right)} \le CD \left(\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H \right). \tag{44}$$

The additive and gap-dependent noise terms are direct geometric sums:

$$\sum_{s=0}^{S-1} \left(\frac{\sigma_0 D}{\Delta_s} \right)^q = (\sigma_0 D)^q \Delta_0^{-q} \frac{2^{qS} - 1}{2^q - 1} \le \frac{2^q}{2^q - 1} \left(\frac{\sigma_0 D}{\varepsilon} \right)^q, \tag{45}$$

$$\sum_{s=0}^{S-1} \left(\frac{\sigma_2 D}{\sqrt{\Delta_s}} \right)^q = (\sigma_2 D)^q \Delta_0^{-q/2} \frac{2^{qS/2} - 1}{2^{q/2} - 1} \le \frac{2^{q/2}}{2^{q/2} - 1} \left(\frac{\sigma_2 D}{\sqrt{\varepsilon}} \right)^q. \tag{46}$$

Both last inequalities use $2^S < 2\Delta_0/\varepsilon$. Thus the constants depend on p through q, but not on Δ_0 or ε. For the gradient-dependent term, apply Lemma 12 with $r = q/2 \ge 1$:

$$\begin{aligned} \sum_{s=0}^{S-1} \left[\sigma_1 D \sqrt{C_\star \left(\frac{H_0}{\Delta_s} + H_1 \right)} \right]^q &\le C_p (\sigma_1 D)^q \left[\left(\frac{H_0}{\varepsilon} \right)^{q/2} H_1^{q/2} J_H \right] \\ &\le C_p \left(\sigma_1 D \left[\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H^{1/q} \right] \right)^q. \end{aligned} \tag{47}$$

The last inequality uses $x^q + y^q \le (x + y)^q$ for $x, y \ge 0$ with $x = \sqrt{H_0/\varepsilon}$ and $y = \sqrt{H_1}\, J_H^{1/q}$. Combining the deterministic sum equation 44, the additive sum equation 45, the gap-dependent sum equation 46, and the gradient-dependent sum equation 47 with the restart event above gives exactly the oracle-complexity bound and the success probability stated in Theorem 9. This sharper accounting changes only the restart summation; the one-phase theorem is unchanged.

Proof of Theorem 10. On a successful phase, $F(w_s) \le \Delta_s$ and strong convexity imply

$$\frac{\mu}{2}\|w_s - x^\star\|^2 \le F(w_s) \le \Delta_s,$$

and hence $\|w_s - x^\star\| \le \sqrt{2\Delta_s/\mu}$. Set

$$Q_s = B\left(w_s, \sqrt{\frac{2\Delta_s}{\mu}}\right), \qquad D_s = 2\sqrt{\frac{2\Delta_s}{\mu}}.$$

Then $w_s, x^\star \in Q_s$, and D_s is the diameter of Q_s, as required by Theorem 8. Substitution into the deterministic phase term gives

$$D_s\sqrt{C_\star\left(\frac{H_0}{\Delta_s} + H_1\right)} = 2\sqrt{\frac{2C_\star}{\mu}}\sqrt{H_0 + H_1\Delta_s} \le 2\sqrt{\frac{2C_\star}{\mu}}\left(\sqrt{H_0} + \sqrt{H_1\Delta_s}\right).$$

The first part is constant across phases and therefore costs S. For the second part, substituting $\Delta_s = 2^{-s}\Delta_0$ gives

$$\sum_{s=0}^{S-1}\sqrt{\Delta_s} = \sqrt{\Delta_0}\frac{1 - 2^{-S/2}}{1 - 2^{-1/2}} \le \frac{\sqrt{\Delta_0}}{1 - 2^{-1/2}}.$$

Thus the summed deterministic contribution is

$$O\left(\sqrt{\frac{H_0}{\mu}}\,S + \sqrt{\frac{H_1\Delta_0}{\mu}}\right),$$

up to the absolute factor depending on $C_\star$.

For additive noise,

$$\left(\frac{\sigma_0 D_s}{\Delta_s}\right)^q = \left(\frac{2\sqrt{2}\,\sigma_0}{\sqrt{\mu\Delta_s}}\right)^q,$$

and substituting $\Delta_s = 2^{-s}\Delta_0$ gives

$$\sum_{s=0}^{S-1}\left(\frac{\sigma_0 D_s}{\Delta_s}\right)^q = \left(\frac{2\sqrt{2}\,\sigma_0}{\sqrt{\mu}}\right)^q \Delta_0^{-q/2}\frac{2^{qS/2} - 1}{2^{q/2} - 1} \le C_p\left(\frac{\sigma_0}{\sqrt{\mu\varepsilon}}\right)^q.$$

Here the last inequality again uses $2^S < 2\Delta_0/\varepsilon$. For gap-dependent noise,

$$\left(\frac{\sigma_2 D_s}{\sqrt{\Delta_s}}\right)^q = \left(\frac{2\sqrt{2}\,\sigma_2}{\sqrt{\mu}}\right)^q$$

is constant across phases. Summing exactly S identical terms therefore contributes

$$\left(\frac{2\sqrt{2}\,\sigma_2}{\sqrt{\mu}}\right)^q S.$$

For the gradient-dependent contribution, substitution of D_s gives

$$\left[\sigma_1 D_s\sqrt{C_\star\left(\frac{H_0}{\Delta_s} + H_1\right)}\right]^q = \left(\frac{2\sqrt{2C_\star}\,\sigma_1}{\sqrt{\mu}}\right)^q (H_0 + H_1\Delta_s)^{q/2}.$$

Since $q/2 \ge 1$, $(x+y)^{q/2} \le 2^{q/2-1}(x^{q/2} + y^{q/2})$. Therefore

$$\sum_{s=0}^{S-1}\Delta_s^{q/2} = \Delta_0^{q/2}\frac{1 - 2^{-qS/2}}{1 - 2^{-q/2}} \le \frac{\Delta_0^{q/2}}{1 - 2^{-q/2}},$$

while $\sum_{s=0}^{S-1} H_0^{q/2} = H_0^{q/2} S$. Using these two identities,

$$\sum_{s=0}^{S-1} \left[\sigma_1 D_s \sqrt{C_\star \left(\frac{H_0}{\Delta_s} + H_1\right)}\right]^q \le C_p \left(\frac{\sigma_1}{\sqrt{\mu}}\right)^q \left(H_0^{q/2} S + (H_1 \Delta_0)^{q/2}\right)$$
$$\le C_p \left(\frac{\sigma_1}{\sqrt{\mu}} \left[\sqrt{H_0}\, S^{1/q} + \sqrt{H_1 \Delta_0}\right]\right)^q. \tag{48}$$

The last line again uses $x^q + y^q \le (x+y)^q$. Equation 48 is the gradient-dependent contribution in Theorem 10. Adding the constant phase cost S, the deterministic contribution, and the three stochastic sums proves the complexity stated there. The restart-event argument gives success probability at least $1-\rho$. Finally, if $\sigma_0 = 0$, every remaining phase contribution is either constant in s or summable from the initial phase, which explains the logarithmic target-accuracy dependence emphasized in the main text.

D Power-Law Noise and Finite-Sum Derivation

D.1 Power-law phase diagram

Localized phase scale. We prove Theorem 11 by isolating the only part of the one-phase cost that changes under equation 7. On the localized trajectory, Lemma 5 gives $F(u_t) \le 4\Delta$. Hence

$$\mathbb{E}_t \|g(u_t, \xi_t) - \nabla f(u_t)\|^p \le \tau_\alpha^p (4\Delta)^\alpha,$$

so the phase moment scale may be chosen as

$$\varsigma_{p,\Delta} = 4^{\alpha/p} \tau_\alpha \Delta^{\alpha/p}.$$

Substitution into the stochastic term of equation 26 gives

$$\left(\frac{\varsigma_{p,\Delta} D}{\Delta}\right)^q = 4^{\alpha q/p} (\tau_\alpha D)^q \Delta^{-q+\alpha q/p}.$$

Because $q = p/(p-1)$ and $q/p = 1/(p-1)$,

$$-q + \frac{\alpha q}{p} = \frac{\alpha - p}{p-1}.$$

Therefore, in the convex case, the stochastic cost of phase s is, up to constants depending only on p and α,

$$(\tau_\alpha D)^q \Delta_s^{(\alpha-p)/(p-1)}. \tag{49}$$

Convex restart. There are three cases:

- If $\alpha < p$, set $a = (p-\alpha)/(p-1) > 0$. Substituting $\Delta_s = 2^{-s}\Delta_0$ gives
$$\sum_{s=0}^{S-1} \Delta_s^{-a} = \Delta_0^{-a} \frac{2^{aS} - 1}{2^a - 1} \le \frac{2^a}{2^a - 1} \varepsilon^{-a}.$$
Thus the last phases determine the order $\varepsilon^{-(p-\alpha)/(p-1)}$.
- If $\alpha = p$, the exponent is zero. Every phase has the same stochastic cost, so the sum contributes $S = O(\log(\Delta_0/\varepsilon))$.
- If $\alpha > p$, set $a = (\alpha - p)/(p-1) > 0$. Then
$$\sum_{s=0}^{S-1} \Delta_s^a = \Delta_0^a \frac{1 - 2^{-aS}}{1 - 2^{-a}} \le \frac{\Delta_0^a}{1 - 2^{-a}}.$$
Thus the initial phases determine the order $\Delta_0^{(\alpha-p)/(p-1)}$.

These are exactly the three branches of $\mathcal{K}_p(\alpha)$ in Theorem 11.

Strongly convex restart. Under strong convexity, the phase diameter is $D_s = 2\sqrt{2\Delta_s/\mu}$. Consequently,

$$D_s^q = 2^{3q/2}\mu^{-q/2}\Delta_s^{q/2},$$

and multiplying equation 49 by the additional factor $\Delta_s^{q/2}/(\mu^{q/2}D^q)$ changes the exponent to

$$\frac{\alpha - p}{p-1} + \frac{q}{2} = \frac{\alpha - p/2}{p-1}.$$

Thus the strongly convex phase cost is

$$C_{p,\alpha}\left(\frac{\tau_\alpha}{\sqrt{\mu}}\right)^q \Delta_s^{(\alpha-p/2)/(p-1)}. \tag{50}$$

For $\alpha < p/2$, substituting $a = (p/2 - \alpha)/(p-1)$ into equation 41 gives

$$\sum_{s=0}^{S-1} \Delta_s^{(\alpha-p/2)/(p-1)} \le \frac{2^a}{2^a - 1}\varepsilon^{-a}.$$

For $\alpha = p/2$, the exponent is zero and the sum equals S. For $\alpha > p/2$, substituting $a = (\alpha - p/2)/(p-1)$ into equation 42 gives

$$\sum_{s=0}^{S-1} \Delta_s^{(\alpha-p/2)/(p-1)} \le \frac{\Delta_0^a}{1 - 2^{-a}}.$$

These are the three branches of $\mathcal{K}_{p/2}(\alpha)$.

Total complexity. The deterministic part of each phase is unchanged, and the restart-event argument from Appendix C supplies the probability $1 - \rho$. Returning to the claim of Theorem 11, the convex and strongly convex stochastic contributions are therefore exactly the two bounds stated there, which completes the proof.

D.2 Finite-sum interpolation

Component-growth assumptions. Let $f(x) = \mathbb{E}_\xi f_\xi(x)$, $g(x,\xi) = \nabla f_\xi(x)$, and fix $x^\star \in \arg\min f$. Suppose

$$\mathbb{E}\|g(x^\star,\xi)\|^p \le a_0^p, \qquad \mathbb{E}\|g(x,\xi) - g(x^\star,\xi)\|^p \le a_1^p\|\nabla f(x)\|^p + a_2^p F(x)^{p/2}. \tag{51}$$

Then Assumption 3 holds with

$$\sigma_0^p = 2^{p-1}a_0^p, \qquad \sigma_1^p = 2^{2p-1}a_1^p, \qquad \sigma_2^p = 2^{2p-1}a_2^p. \tag{52}$$

Centered decomposition. We verify the reduction directly. Let $A(x,\xi) = g(x,\xi) - g(x^\star,\xi)$ and $B(\xi) = g(x^\star,\xi)$. Since $\mathbb{E}B = \nabla f(x^\star) = 0$ and $\mathbb{E}A = \nabla f(x)$,

$$g(x,\xi) - \nabla f(x) = (A - \mathbb{E}A) + B.$$

Conditional moment bound. The equality $\mathbb{E}B = 0$ uses differentiability of f and first-order optimality at the unconstrained minimizer $x^\star$. For $p \in (1,2]$, $\|u+v\|^p \le 2^{p-1}(\|u\|^p + \|v\|^p)$, so

$$\mathbb{E}\|g(x,\xi) - \nabla f(x)\|^p \le 2^{p-1}\mathbb{E}\|A - \mathbb{E}A\|^p + 2^{p-1}\mathbb{E}\|B\|^p.$$

Again by the same inequality and Jensen's inequality,

$$\mathbb{E}\|A - \mathbb{E}A\|^p \le 2^{p-1}(\mathbb{E}\|A\|^p + \|\mathbb{E}A\|^p) \le 2^p\mathbb{E}\|A\|^p.$$

The second inequality follows because $\|\mathbb{E}A\|^p \le \mathbb{E}\|A\|^p$. Substituting equation 51 yields

$$\mathbb{E}\|g(x,\xi) - \nabla f(x)\|^p \le 2^{p-1}a_0^p + 2^{2p-1}a_1^p\|\nabla f(x)\|^p + 2^{2p-1}a_2^p F(x)^{p/2},$$

which proves equation 52. For a finite-sum oracle sampled independently at every iteration, the same calculation holds conditionally on $\mathcal{F}_{t-1}$ because the query point is then fixed and the new index is fresh. Hence the resulting constants satisfy the conditional formulation of Assumption 3.

Interpolation endpoint. In particular, exact interpolation gives $g(x^\star, \xi) = 0$ almost surely, so $a_0 = \sigma_0 = 0$ under equation 51. Theorem 10 then has logarithmic target-accuracy dependence, even when $p < 2$.

If $\mathbb{E}\|g(x,\xi) - g(x^\star,\xi)\|^p \le L_p^{p/2} F(x)^{p/2}$ and $\mathbb{E}\|g(x^\star,\xi)\|^p \le \sigma_{\star,p}^p$, then $\sigma_1 = 0$, $\sigma_0^p \le 2^{p-1}\sigma_{\star,p}^p$, and $\sigma_2^p \le 2^{2p-1} L_p^{p/2}$. At $p = 2$ this is the familiar component-difference expected-smoothness scale up to convention-dependent constants (Gower et al., 2019); for $p < 2$ it should be read as a finite-p moment analogue rather than as a unique standard definition.

E Lower Bounds on the Smooth Subclass

Throughout this section $H_1 = 0$, hence Assumption 2 is the standard global smoothness upper model with $L = 2H_0$. Let $q = p/(p-1)$. The purpose of restricting to this subclass is logical: a lower bound on a subclass is automatically a lower bound for the full (H_0, H_1) class. We treat deterministic curvature, additive finite-p noise, and gap-dependent noise separately so that each term in Theorems 9 and 10 can be compared with a matching hard instance.

E.1 Deterministic curvature

The standard first-order lower bounds (Nesterov, 2004) imply, in sufficiently high dimension,

$$N = \Omega\left(D\sqrt{\frac{H_0}{\varepsilon}}\right)$$

for convex objectives and

$$N = \Omega\left(\sqrt{\frac{H_0}{\mu}}\log\frac{\Delta_0}{\varepsilon}\right)$$

for μ-strongly convex objectives in the nontrivial condition-number regime. These match the corresponding deterministic parts of Theorems 9–10 when $H_1 = 0$.

E.2 Additive finite-p noise

We use a two-point testing construction in which information about the sign of the minimizer appears only through a rare oracle response. The rarity of that response is chosen to saturate the pth-moment constraint.

For $v \in \{-1, +1\}$, $\delta > 0$, and $\eta \in (0, 1/2]$, define

$$Z_v = \begin{cases} -v\delta/\eta, & \text{with probability } \eta, \\ 0, & \text{with probability } 1-\eta. \end{cases} \tag{53}$$

Then $\mathbb{E}Z_v = -v\delta$. On the rare event, the centered value $Z_v - \mathbb{E}Z_v = Z_v + v\delta$ has magnitude $\delta(1-\eta)/\eta$; on the complementary event, it has magnitude δ. Therefore

$$\mathbb{E}|Z_v + v\delta|^p = \delta^p\left[(1-\eta) + \eta^{1-p}(1-\eta)^p\right] \le 2\delta^p\eta^{1-p}. \tag{54}$$

Choosing

$$\eta = 2^{1/(p-1)}\left(\frac{\delta}{\sigma_0}\right)^q \tag{55}$$

ensures $\mathbb{E}|Z_v + v\delta|^p \le \sigma_0^p$ whenever $\eta \le 1/2$. Indeed,

$$2\delta^p\eta^{1-p} = 2\delta^p\left[2^{1/(p-1)}\left(\frac{\delta}{\sigma_0}\right)^q\right]^{1-p} = \sigma_0^p.$$

Theorem 13 (Convex additive lower bound)**.** *Let $\rho \le 1/4$, $\varepsilon \le H_0 D^2/32$, and suppose the η in equation 55 with $\delta = 16\varepsilon/D$ is at most $1/2$. Any stochastic first-order method that succeeds uniformly with probability $1-\rho$ on the subclass $H_1 = \sigma_1 = \sigma_2 = 0$ requires*

$$T \ge c_p\left(\frac{\sigma_0 D}{\varepsilon}\right)^q \log\frac{1}{2\rho}.$$

Proof. We construct two one-dimensional quadratics indexed by a hidden sign v. We verify their membership in the smooth subclass and the oracle moment condition, show that an accurate point reveals v, and then lower-bound the number of calls needed to observe the rare sign-dependent response.

Hard instances. Take $Q = [-D/2, D/2]$, $x_0 = 0$, $r = D/4$, and

$$f_v(x) = \frac{\lambda}{2}(x - vr)^2, \qquad \lambda = \frac{64\varepsilon}{D^2}, \qquad \delta = \lambda r = \frac{16\varepsilon}{D}.$$

The minimizer is $x_v^\star = vr \in Q$, and

$$F_v(x) = f_v(x) - f_v(x_v^\star) = \frac{\lambda}{2}(x - vr)^2.$$

The Hessian is the scalar λ. Since $\varepsilon \le H_0 D^2/32$,

$$\frac{\lambda}{2} = \frac{32\varepsilon}{D^2} \le H_0,$$

so f_v satisfies Assumption 2 with $H_1 = 0$.

Oracle validity. Use the oracle

$$g_v(x) := \lambda x + Z_v.$$

Because $\delta = \lambda r$,

$$\mathbb{E}g_v(x) = \lambda x - v\delta = \lambda(x - vr) = \nabla f_v(x),$$

and its centered error is $Z_v - \mathbb{E}Z_v = Z_v + v\delta$. Thus equation 54–equation 55 verify Assumption 3 with $\sigma_1 = \sigma_2 = 0$.

Reduction to sign testing. If $vx \le 0$, then x lies on the wrong side of the origin and $|x - vr| \ge r$. Consequently,

$$F_v(x) \ge \frac{\lambda r^2}{2} = 2\varepsilon.$$

In particular, any output satisfying $F_v(x) \le \varepsilon$ must have the same sign as v and therefore identifies the hidden instance.

Now place the uniform prior on $v \in \{-1, +1\}$. Let $\mathcal{N}_T$ be the event that none of the first T oracle calls produces the rare value in equation 53. Since the rare-event indicators are independent of the adaptive query points,

$$\mathbb{P}(\mathcal{N}_T) = (1 - \eta)^T.$$

On $\mathcal{N}_T$, every response equals λx, regardless of v. Thus the entire adaptive transcript, and hence the algorithm's output distribution, is identical for the two signs. Conditional on $\mathcal{N}_T$, the probability of identifying v is at most $1/2$. The average failure probability is therefore at least $\frac{1}{2}(1 - \eta)^T$, so the worst-case failure probability over the two instances is at least the same quantity.

Number of oracle calls. If the method has failure probability at most ρ on both instances, then $(1 - \eta)^T \le 2\rho$. Taking logarithms gives

$$T \ge \frac{\log(1/(2\rho))}{-\log(1 - \eta)} \ge \frac{1}{2\eta} \log \frac{1}{2\rho},$$

where the last inequality uses $-\log(1 - \eta) \le 2\eta$ for $\eta \le 1/2$. Finally, substituting equation 55 with $\delta = 16\varepsilon/D$ proves the stated lower bound, with a constant depending only on p. □

Theorem 14 (Strongly convex additive lower bound)**.** *Let $\rho \le 1/4$ and $H_0 \ge \mu/2$. Define $\delta = 2\sqrt{\mu\varepsilon}$ and let η be given by equation 55. If $\eta \le 1/2$, any method with uniform success probability $1 - \rho$ on the subclass $H_1 = \sigma_1 = \sigma_2 = 0$ requires*

$$T \ge c_p \left(\frac{\sigma_0}{\sqrt{\mu\varepsilon}}\right)^q \log \frac{1}{2\rho}.$$

Proof. We reuse the rare-signal oracle with quadratic curvature μ. The proof verifies strong convexity and oracle validity, reduces optimization to sign testing, and substitutes the strong-convexity scale into the testing bound.

Hard instances and oracle validity. Use

$$f_v(x) = \frac{\mu}{2}(x - vr)^2, \qquad r = 2\sqrt{\frac{\varepsilon}{\mu}},$$

so $\delta = \mu r = 2\sqrt{\mu\varepsilon}$. The functions are μ-strongly convex, and their quadratic upper model has coefficient $\mu/2 \le H_0$. Thus they belong to the asserted subclass. If $vx \le 0$, then $F_v(x) \ge \mu r^2/2 = 2\varepsilon$, so an ε-accurate output again identifies v. With the oracle $g_v(x) = \mu x + Z_v$, the unbiasedness and moment calculation use equation 54–equation 55, with λ replaced by μ.

Testing reduction and final complexity. On the event with no rare response, the transcript is independent of v, exactly as in the convex construction. Therefore the sign-testing calculation gives $T \ge [2\eta]^{-1}\log(1/(2\rho))$. Substitution of $\delta = 2\sqrt{\mu\varepsilon}$ into equation 55 then gives $\eta^{-1} = \Theta_p((\sigma_0/\sqrt{\mu\varepsilon})^q)$ and proves the claim. □

E.3 Gap/interpolation noise

Set $\sigma_0 = \sigma_1 = 0$ and use Bernoulli thinning of the exact gradient:

$$g(x) = \frac{B}{\eta}\nabla f(x), \qquad B \sim \mathrm{Ber}(\eta). \tag{56}$$

This oracle is unbiased because $\mathbb{E}B = \eta$. Its centered error equals $(B/\eta - 1)\nabla f(x)$. If $\eta \le 1/2$, direct evaluation of the two Bernoulli outcomes gives

$$\begin{aligned}\mathbb{E}_t\|g(x) - \nabla f(x)\|^p &= \left[(1-\eta) + \eta^{1-p}(1-\eta)^p\right]\|\nabla f(x)\|^p \\ &\le 2\eta^{1-p}\|\nabla f(x)\|^p.\end{aligned} \tag{57}$$

Theorem 15 (Convex gap-noise lower bound)**.** *Let $\rho \le 1/4$, $\varepsilon \le H_0 D^2/32$, and set $\lambda = 64\varepsilon/D^2$. Choose η so that*

$$2\eta^{1-p}(2\lambda)^{p/2} = \sigma_2^p,$$

and assume the resulting η is at most $1/2$. Any method with uniform success probability $1-\rho$ on the subclass $H_1 = \sigma_0 = \sigma_1 = 0$ requires

$$T \ge c_p\left(\frac{\sigma_2 D}{\sqrt{\varepsilon}}\right)^q \log\frac{1}{2\rho}.$$

Proof. We use the convex quadratic pair from Theorem 13, replace the additive rare signal by Bernoulli thinning, verify the gap-dependent conditional moment bound, and then repeat the sign-testing reduction with its probability written explicitly.

Conditional moment bound. Use the same two quadratics as in Theorem 13. For a quadratic, $|\nabla f_v(x)|^2 = 2\lambda F_v(x)$, and hence

$$|\nabla f_v(x)|^p = (2\lambda F_v(x))^{p/2}.$$

Therefore equation 57 and the definition of η give

$$\mathbb{E}_t|g(x) - \nabla f_v(x)|^p \le \sigma_2^p F_v(x)^{p/2}.$$

Solving for η yields

$$\eta^{-1} = \Theta_p\left(\frac{\sigma_2}{\sqrt{\lambda}}\right)^q = \Theta_p\left(\frac{\sigma_2 D}{\sqrt{\varepsilon}}\right)^q.$$

Testing reduction. If $B_t = 0$ at every call, every oracle response is zero, independently of the sign v. The full adaptive transcript then contains no information about the minimizer. Since an ε-accurate

output must identify v, the same calculation as in Theorem 13 gives the following explicit testing inequality:

$$\rho \geq \frac{1}{2}\mathbb{P}\{B_1 = \cdots = B_T = 0\} = \frac{1}{2}(1-\eta)^T.$$

Taking logarithms gives $T\log(1/(1-\eta)) \geq \log(1/(2\rho))$. Because $\eta \leq 1/2$ implies $\log(1/(1-\eta)) \leq \eta/(1-\eta) \leq 2\eta$, we conclude that $T \geq [2\eta]^{-1}\log(1/(2\rho))$. Combining this inequality with the displayed value of η^{-1} proves the claim. □

Lemma 16 (Thinning transfers exact-gradient lower bounds)**.** *Suppose a hard distribution over functions has the property that every exact first-order method using fewer than K informative gradients has average failure at least $c_0 > 0$. Replacing the exact oracle by equation 56, any stochastic method with constant success probability bounded away from $1 - c_0$ needs $T = \Omega(K/\eta)$ calls.*

Proof. Let S_T denote the number of nonzero thinned responses among T calls. We first couple the stochastic transcript to an exact-gradient transcript containing only S_T informative queries, then control S_T and invoke the assumed exact-gradient lower bound.

Transcript simulation. The number of nonzero answers is $S_T := \sum_{t=1}^{T} B_t \sim \mathrm{Binomial}(T, \eta)$, even though the query points may be adaptive, because the Bernoulli variables are fresh and independent. Conditional on the Bernoulli sequence, every nonzero response is $\nabla f(x_t)/\eta$ and therefore reveals exactly the same information as an exact gradient, while a zero response contains no instance-dependent information. Thus the stochastic transcript can be simulated by an exact-oracle method making only S_T informative queries.

Informative-query count. If $T < K/(2\eta)$, then $\mathbb{E}S_T = T\eta < K/2$, and Markov's inequality gives $\mathbb{P}\{S_T \geq K\} \leq 1/2$. On the event $\{S_T < K\}$, the assumed exact-oracle lower bound gives conditional average failure at least c_0. Hence the unconditional average failure is at least $c_0/2$. This proves that a constant success probability bounded away from $1 - c_0/2$ requires $T = \Omega(K/\eta)$. Yao's minimax principle converts the distributional statement into a randomized worst-case lower bound. □

Theorem 17 (Strongly convex gap-noise lower bound)**.** *There is a universal $c > 0$ such that, on a fixed-condition-number smooth strongly convex subclass (for example, $L = 4\mu$, corresponding to $H_0 = 2\mu$), and in the regime $\sigma_2 \gtrsim_p \sqrt{\mu}$, any stochastic first-order method with constant success probability requires*

$$T \geq c_p \left(\frac{\sigma_2}{\sqrt{\mu}}\right)^q \log\frac{\Delta_0}{\varepsilon}.$$

For arbitrary confidence, an additional term of order $\eta^{-1}\log(1/\rho)$ is unavoidable; no claim is made that all upper and lower confidence logarithms coincide.

Proof. We start from a fixed-condition-number deterministic zero-chain requiring K informative gradients. We then verify that Bernoulli thinning satisfies the gap-dependent moment model, apply Lemma 16, and finally account for the confidence-dependent probability of receiving no informative response.

Deterministic hard family. Take an L-smooth, μ-strongly convex zero-chain family with $L = 4\mu$ from the first-order lower-bound construction of Nesterov (2004). For this fixed condition number, the construction states that, in sufficiently high dimension, every exact first-order method needs

$$K = \Omega\left(\log\frac{\Delta_0}{\varepsilon}\right)$$

informative gradients to reduce the initial gap Δ_0 to ε. Because our convention is $L = 2H_0$, choosing $H_0 = 2\mu$ places the family in the subclass of Assumption 2.

Conditional moment bound for the thinned oracle. For every L-smooth convex function, apply its global upper model at $x - \nabla f(x)/L$. Comparing the resulting value with $f^\star$ gives $\|\nabla f(x)\|^2 \leq 2LF(x)$. Substituting this inequality into equation 57 yields

$$\mathbb{E}_t\|g(x_t) - \nabla f(x_t)\|^p \leq 2\eta^{1-p}(2L)^{p/2}F(x_t)^{p/2}.$$

Choose η by

$$2\eta^{1-p}(2L)^{p/2} = \sigma_2^p.$$

The assumption $\sigma_2 \gtrsim_p \sqrt{\mu}$ ensures that the resulting η lies in $(0, 1/2]$. Since $L = 4\mu$,

$$\eta^{-1} = \Theta_p\left(\frac{\sigma_2}{\sqrt{L}}\right)^q = \Theta_p\left(\frac{\sigma_2}{\sqrt{\mu}}\right)^q.$$

Transfer and confidence. Lemma 16 now gives $T = \Omega(K/\eta)$ and proves the constant-confidence term. For confidence $1-\rho$, even a two-instance test cannot succeed before at least one informative Bernoulli outcome appears; the probability of no such outcome is $(1-\eta)^T$. Repeating the logarithmic calculation from Theorem 13 yields the additional necessary term $\Omega(\eta^{-1}\log(1/\rho))$. □

F Multiplicative Noise: Universality versus Specialized Acceleration

Set $H_1 = \sigma_0 = \sigma_2 = 0$ and suppose

$$\mathbb{E}_t\|g(x_t) - \nabla f(x_t)\|^p \le \sigma_1^p\|\nabla f(x_t)\|^p. \tag{58}$$

The thinned oracle equation 56 satisfies this condition when $2\eta^{1-p} \le \sigma_1^p$, so the transfer lemma gives the information-theoretic benchmark

$$T = \Omega_p\big((1+\sigma_1^q)K\big) \tag{59}$$

Equation 59 is the information-theoretic benchmark against which we compare the universal phase construction below. Here K denotes any deterministic exact-gradient lower complexity at constant confidence.

The universal phase proof instead bounds $\|\nabla f\|$ by a phase constant before robustification; its sample cost can then contain $K + (\sigma_1 K)^q$. Averaging b stochastic gradients does not remove this term within the same argument. A finite-p batch mean decreases the relative scale as $b^{-1/q}$, so the number of accelerated outer iterations becomes $K + \sigma_1^q K^q/b$, while raw samples equal

$$bK + \sigma_1^q K^q. \tag{60}$$

Equation 60 shows that the problematic term is independent of b.

At $p = 2$, specialized AGNES dynamics achieves $O((1+\sigma_1^2)K)$ in expectation for smooth convex and strongly convex minimization (Gupta et al., 2024), showing that the benchmark is algorithmically meaningful. For $1 < p < 2$, robust mean estimation gives the following partial reduction.

Lemma 18 (Robust relative gradient estimator)**.** *At a predictable point x, take m fresh samples satisfying equation 58. There is a geometric median-of-means estimator $\widehat{g}(x)$ such that with probability at least $1-\delta$,*

$$\|\widehat{g}(x) - \nabla f(x)\| \le K_p\sigma_1\left(\frac{\log(2/\delta)}{m}\right)^{1/q}\|\nabla f(x)\|.$$

Thus constant relative accuracy uses $m = O_p(\sigma_1^q\log(1/\delta))$ samples.

Proof. Conditionally on the history, we identify the finite-p moment scale of the fresh stochastic gradients and apply the geometric median-of-means lemma. We then rewrite its exponent as $1/q$ and solve the resulting inequality for a prescribed relative accuracy.

Conditional reduction. Condition on the history that determines the predictable point x. The fresh stochastic gradients are then independent random vectors with conditional mean $\nabla f(x)$ and conditional pth central moment at most

$$\nu^p := \sigma_1^p\|\nabla f(x)\|^p.$$

Apply Lemma 19 below with this value of ν and $n = m$. It gives

$$\|\widehat{g}(x) - \nabla f(x)\| \le K_p\nu\left(\frac{\log(2/\delta)}{m}\right)^{(p-1)/p}.$$

Since $(p-1)/p = 1/q$ and $\nu = \sigma_1\|\nabla f(x)\|$, this is the claimed inequality. To make the relative error at most any fixed constant $c \in (0,1)$, it is enough to take $m \ge C_{p,c}\sigma_1^q\log(2/\delta)$. □

This shows that the heavy-tailed mean-estimation layer itself has the desired σ_1^q scaling. What remains is a high-probability accelerated outer method that is robust to the resulting constant relative error without adding unfavorable conditioning requirements. Existing relative-inexact acceleration analyses impose additional restrictions (Vasin et al., 2023; Kornilov et al., 2025). We therefore do not state a general finite-p matching theorem for σ_1.

G Gradient-Only End-to-End Initialization

The restart statements use a supplied upper bound $F(x_0) \le \Delta_0$. This can be obtained from stochastic gradients without evaluating f, $f^\star$, or an exact gradient.

G.1 Choosing the auxiliary set

The convex result requires a closed bounded convex set Q with $x_0, x^\star \in Q$ and a tractable Euclidean projection; this set controls the iterates but does not constrain the original problem. One may take a known compact convex parameter set containing x_0 and an optimizer, or $Q = B(x_0, R)$ when a radius $R \ge \|x_0 - x^\star\|$ is available, in which case $D = 2R$. Regularization can provide such a radius: if $x_0 = 0$ and $f(x) = \ell(x) + (\lambda/2)\|x\|^2$ for a nonnegative convex loss ℓ, then $f(x^\star) \le f(0)$ gives $\|x^\star\| \le \sqrt{2\ell(0)/\lambda}$. Here f is also λ-strongly convex, so the shrinking-ball construction of Appendix C applies without prior knowledge of $x^\star$.

G.2 Robust gradient estimation

Lemma 19 (Finite-p geometric median-of-means)**.** *Let X_i be independent, $\mathbb{E}X_i = \mu$, and $\mathbb{E}\|X_i - \mu\|^p \le \nu^p$ for $1 < p \le 2$. For $n \ge c\log(2/\delta)$, a geometric median-of-means estimator $\widehat{\mu}$ satisfies with probability at least $1 - \delta$,*

$$\|\widehat{\mu} - \mu\| \le K_p \nu \left(\frac{\log(2/\delta)}{n}\right)^{(p-1)/p}.$$

Proof. We divide the n independent observations into independent blocks, control each block mean by a finite-p moment inequality and Markov's inequality, amplify the fraction of good blocks by a Chernoff bound, and finally use geometric-median stability to obtain the stated confidence bound.

Block means. Let $k = \lceil c_0 \log(2/\delta) \rceil$ and split the observations into k disjoint blocks of common size $b = \lfloor n/k \rfloor$, discarding at most $k - 1$ observations. For block j, let

$$\bar{X}_j := \frac{1}{b} \sum_{i \in \mathcal{I}_j} X_i.$$

The finite-p moment inequality for sums of independent mean-zero random vectors in a Hilbert space gives

$$\mathbb{E}\|\bar{X}_j - \mu\|^p \le C_p \nu^p b^{1-p}.$$

This is the vector analogue of the von Bahr–Esseen inequality; see, for example, Minsker (2015). Markov's inequality therefore implies

$$\mathbb{P}\left\{\|\bar{X}_j - \mu\| > (4C_p)^{1/p} \nu b^{-(p-1)/p}\right\} \le \frac{1}{4}.$$

Probability amplification. The blocks are independent. A Chernoff bound shows that, for a sufficiently large absolute choice of c_0, with probability at least $1 - \delta$ more than half of the block means lie in the ball centered at μ with radius

$$r_p := (4C_p)^{1/p} \nu b^{-(p-1)/p}.$$

Geometric-median aggregation. On this event, the geometric-median stability lemma of Minsker (2015) implies that any geometric median of the block means is at distance at most a universal constant times r_p from μ. Since $b \ge c_1 n/k$ when $n \ge ck$, we conclude that

$$\|\widehat{\mu} - \mu\| \le K_p \nu \left(\frac{k}{n}\right)^{(p-1)/p} \le K_p \nu \left(\frac{\log(2/\delta)}{n}\right)^{(p-1)/p}.$$

All constants depending on the Hilbert-space moment inequality and on the geometric-median stability factor are absorbed into K_p. □

Let $\kappa = K_p(\log(2/\delta)/n_0)^{(p-1)/p}$ and $\widehat{r}_0 = \|\widehat{g}_0\|$.

G.3 Convex initialization

Assume the auxiliary set Q is already known and contains both x_0 and $x^\star$. Define $r_0 := \|\nabla f(x_0)\|$.

Gap bound from convexity. Convexity at x_0 gives

$$F(x_0) = f(x_0) - f(x^\star) \le \langle \nabla f(x_0), x_0 - x^\star \rangle \le r_0 \|x_0 - x^\star\| \le D r_0.$$

Conditional moment bound. Consequently, Assumption 3 at x_0 gives the pth-moment scale

$$\left(\sigma_0^p + \sigma_1^p r_0^p + \sigma_2^p F(x_0)^{p/2}\right)^{1/p} \le \left(\sigma_0^p + \sigma_1^p r_0^p + \sigma_2^p (D r_0)^{p/2}\right)^{1/p} \le \sigma_0 + \sigma_1 r_0 + \sigma_2 \sqrt{D r_0},$$

where the last step is the triangle inequality for the ℓ_p norm of three nonnegative scalars. On the good event from Lemma 19,

$$\|\widehat{g}_0 - \nabla f(x_0)\| \le \kappa \left(\sigma_0 + \sigma_1 r_0 + \sigma_2 \sqrt{D r_0}\right).$$

Observable upper bound. The triangle inequality $r_0 \le \|\widehat{g}_0\| + \|\widehat{g}_0 - \nabla f(x_0)\|$ therefore yields

$$r_0 \le \widehat{r}_0 + \kappa\sigma_0 + \kappa\sigma_1 r_0 + \kappa\sigma_2 \sqrt{D r_0}.$$

Quadratic inequality. If $\kappa\sigma_1 < 1$, define $\alpha = 1 - \kappa\sigma_1$, $\beta = \kappa\sigma_2\sqrt{D}$, and $\gamma = \widehat{r}_0 + \kappa\sigma_0$. With $y = \sqrt{r_0}$, $\alpha y^2 - \beta y - \gamma \le 0$. Since $\alpha > 0$ and $y \ge 0$, the quadratic inequality places y below its nonnegative root:

$$y \le \frac{\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}.$$

Squaring and using $F(x_0) \le D r_0$ gives

$$r_0 \le \bar{r}_0^{\mathrm{cvx}} := \left(\frac{\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}\right)^2, \qquad \widehat{\Delta}_0^{\mathrm{cvx}} := D\bar{r}_0^{\mathrm{cvx}}. \tag{61}$$

Thus equation 61 makes $\widehat{\Delta}_0^{\mathrm{cvx}}$ a computable upper bound on $F(x_0)$ on the geometric-median-of-means event. It suffices to choose n_0 such that $\kappa\sigma_1 \le 1/2$, an accuracy-independent condition that keeps the denominator uniformly away from zero.

G.4 Strongly convex initialization

Gap bound from strong convexity. For a differentiable μ-strongly convex function, Assumption 1 with $x = x_0$ and $y = x^\star$ gives

$$F(x_0) \le \langle \nabla f(x_0), x_0 - x^\star \rangle - \frac{\mu}{2}\|x_0 - x^\star\|^2.$$

Maximizing the right-hand side over $s = \|x_0 - x^\star\| \ge 0$ and using Cauchy–Schwarz yields

$$F(x_0) \le \max_{s \ge 0} \left(r_0 s - \frac{\mu}{2} s^2\right) = \frac{r_0^2}{2\mu}.$$

Effective relative noise. Define the effective relative coefficient

$$\bar{\sigma}_{1,\mu} := \left(\sigma_1^p + \frac{\sigma_2^p}{(2\mu)^{p/2}}\right)^{1/p}. \tag{62}$$

The definition in equation 62 combines the two oracle-moment terms that are proportional to r_0^p. The oracle moment at x_0 is then bounded by

$$\begin{aligned}\sigma_0^p + \sigma_1^p r_0^p + \sigma_2^p F(x_0)^{p/2} &\le \sigma_0^p + \left(\sigma_1^p + \frac{\sigma_2^p}{(2\mu)^{p/2}}\right) r_0^p \\ &= \sigma_0^p + \bar{\sigma}_{1,\mu}^p r_0^p.\end{aligned}$$

Taking the pth root, applying Lemma 19, and using the triangle inequality gives

$$r_0 \le \widehat{r}_0 + \kappa\sigma_0 + \kappa\bar{\sigma}_{1,\mu} r_0.$$

Observable upper bound. If $\kappa\bar{\sigma}_{1,\mu} < 1$, rearrangement yields

$$\bar{r}_0^{\mathrm{sc}} := \frac{\widehat{r}_0 + \kappa\sigma_0}{1 - \kappa\bar{\sigma}_{1,\mu}}, \qquad \widehat{\Delta}_0^{\mathrm{sc}} := \frac{(\bar{r}_0^{\mathrm{sc}})^2}{2\mu}. \tag{63}$$

Equation 63 and the first inequality above then give $F(x_0) \le \widehat{\Delta}_0^{\mathrm{sc}}$ on the geometric-median-of-means event. Splitting the total failure probability, for example assigning $\rho/2$ to initialization and $\rho/2$ to the restart run, gives an end-to-end success probability at least $1 - \rho$. The construction uses sampled stochastic gradients and known upper bounds on the model parameters, but no objective values, value of $f^\star$, or exact gradients.

H Endpoints and Special Cases

H.1 $p = 2$

When $p = 2$, the conjugate exponent is $q = 2$. Assumption 3 becomes

$$\mathbb{E}_t\|g(x,\xi_t) - \nabla f(x)\|^2 \le \sigma_0^2 + \sigma_1^2\|\nabla f(x)\|^2 + \sigma_2^2 F(x).$$

No step in Lemma 6 becomes singular: its clipping-bias bound is $2\varsigma_{2,\Delta}^2/\Lambda$, and its conditional second-moment bound is $2\varsigma_{2,\Delta}^2$.

Convex specialization. Substituting $q = 2$ into Theorem 9 gives

$$\widetilde{O}_\rho\left[\begin{aligned} &S + D\left(\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H\right) + \left(\frac{\sigma_0 D}{\varepsilon}\right)^2 \\ &\quad + \left(\sigma_1 D\left[\sqrt{\frac{H_0}{\varepsilon}} + \sqrt{H_1}\, J_H^{1/2}\right]\right)^2 + \left(\frac{\sigma_2 D}{\sqrt{\varepsilon}}\right)^2 \end{aligned}\right].$$

Thus the three pure stochastic accuracy scales are ε^{-2} for residual noise and ε^{-1} for the gradient- and gap-dependent components, while the displayed square retains the H_0–H_1 cross term.

Strongly convex specialization. Theorem 10 specializes to

$$\widetilde{O}_\rho\left[\left(1 + \sqrt{\frac{H_0}{\mu}}\right) S + \sqrt{\frac{H_1\Delta_0}{\mu}} + \frac{\sigma_0^2}{\mu\varepsilon} + \frac{\sigma_1^2}{\mu}\left(\sqrt{H_0 S} + \sqrt{H_1\Delta_0}\right)^2 + \frac{\sigma_2^2}{\mu} S\right].$$

Only the σ_0 term is polynomial in the target accuracy; the terms multiplied by S are logarithmic because $S = \lceil \log_2(\Delta_0/\varepsilon)\rceil$.

H.2 NORM-SUB-GAUSSIAN ENDPOINT

If instead the centered oracle is conditionally norm-sub-Gaussian with a state-dependent proxy comparable to $\sigma_0^2 + \sigma_1^2\|\nabla f(x)\|^2 + \sigma_2^2 F(x)$, localization again creates the phase proxy $\nu_{\mathrm{SG},\Delta}^2 = \sigma_0^2 + \sigma_1^2 G_\Delta^2 + 4\sigma_2^2\Delta$. The finite-$p$ clipping proof remains applicable with $p = 2$ and already gives the accuracy exponents displayed above. An unclipped mini-batch alternative would require a separately stated vector sub-Gaussian mean-concentration inequality, including its precise definition of the proxy and its numerical constants. Because those data are not part of the present assumptions, no theorem in this paper relies on that alternative; we record it only as a possible implementation endpoint.

H.3 DETERMINISTIC LIMIT AND OTHER AUDIT CASES

No stochastic noise. If $\sigma_0 = \sigma_1 = \sigma_2 = 0$, then $g(x, \xi_t) = \nabla f(x)$ almost surely under Assumption 3. Hence $\zeta_t = \beta_t = 0$, the two martingale channels vanish, and equation 6 reduces to $\Phi_t \le \Phi_{t-1}$. The restart sums retain only the constant and deterministic curvature terms.

No residual noise. If $\sigma_0 = 0$, the additive term in equation 40 disappears. In the strongly convex case, the remaining σ_1 contribution is the sum of a constant per phase and a decreasing geometric sequence, while the σ_2 contribution is constant per phase. Thus the target-accuracy dependence is logarithmic through S, with no term proportional to a positive power of $1/\varepsilon$.

No gradient-dependent noise. If $\sigma_1 = 0$, the gradient-dependent term is absent from $\varsigma_{p,\Delta}$. The crossover calculation with exponent $q/2$ is then unnecessary; the stochastic complexity contains only the additive and gap-dependent geometric sums.

No gap-dependent noise. If $\sigma_2 = 0$, the term $(\sigma_2 D/\sqrt{\Delta})^q$ vanishes in every phase. The analysis and clipping threshold remain valid without modification because all inequalities were upper bounds by sums of nonnegative noise components.

Globally smooth endpoint. If $H_1 = 0$, then $r_H = \infty$, $C_\star = 2$, and $\mathcal{H}_\Delta = H_0$. Every local-model radius check becomes automatic, $J_H = 0$, and the deterministic part reduces to the classical accelerated dependence $D\sqrt{H_0/\varepsilon}$ in the convex case and $\sqrt{H_0/\mu}\,S$ in the strongly convex case. Since our convention is $L = 2H_0$, this is the usual globally L-smooth subclass up to the fixed factor $\sqrt{2}$.

These reductions verify that the unified formulas remain well-defined at every endpoint and that no proof step divides by a parameter that may be zero.

I HORIZON-DRIVEN PARAMETER-FREE INNER METHOD

This section removes $p, \sigma_0, \sigma_1, \sigma_2, H_0, H_1$ from the *runtime inputs* of a phase. They remain in an a posteriori sufficient horizon condition. The phase still receives $(w, \Delta, Q, D, T, \rho)$, and the restart layer still needs a valid phase level; strong convexity uses μ to form shrinking balls.

I.1 A RADIUS-AWARE SELF-BOUND

For analysis only, define

$$\Gamma_\Delta := H_0 + 4H_1\Delta + \frac{4\Delta}{r_H^2}, \tag{64}$$

with $r_H^{-2} = 0$ if $r_H = \infty$. The quantity in equation 64 upper-bounds every curvature and radius term that appears after localization to $F(x) \le 4\Delta$.

Lemma 20 (Radius-aware self-bound)**.** *For every x,*

$$\|\nabla f(x)\|^2 \le 4F(x)\left(H_0 + H_1 F(x) + \frac{F(x)}{r_H^2}\right).$$

Proof. Fix x and abbreviate $g := \|\nabla f(x)\|$ and $H_x := \mathcal{H}(x)$. We compare the unconstrained quadratic-model step with the local-model radius, use the shorter of the two admissible steps, and then handle the zero-curvature cases separately. If $g = 0$, the claim is immediate. Suppose first that $H_x > 0$ and the unconstrained quadratic-model step lies inside the validity radius:

$$\frac{g}{2H_x} \le r_H.$$

Interior quadratic-model step. Set $y = x - \nabla f(x)/(2H_x)$. Then $\|y - x\| = g/(2H_x) \le r_H$, so Assumption 2 applies and gives

$$f(y) \le f(x) - \frac{g^2}{2H_x} + H_x \frac{g^2}{4H_x^2} = f(x) - \frac{g^2}{4H_x}.$$

Since $f^\star \le f(y)$, it follows that $F(x) \ge g^2/(4H_x)$, or

$$g^2 \le 4F(x)H_x.$$

Full-radius step. It remains to treat the case in which the quadratic-model step would be longer than r_H, namely $g/(2H_x) > r_H$. Take the full-radius step

$$y = x - r_H \frac{\nabla f(x)}{g}.$$

The local model and $f^\star \le f(y)$ give

$$F(x) \ge r_H g - H_x r_H^2.$$

The case condition implies $H_x r_H < g/2$, and hence

$$F(x) > \frac{r_H g}{2}, \qquad g^2 < \frac{4F(x)^2}{r_H^2}.$$

The first case is bounded by the first term and the second case by the second term in

$$4F(x)\left(H_x + \frac{F(x)}{r_H^2}\right).$$

This proves the result when $H_x > 0$.

Degenerate curvature cases. If $H_x = 0$ and $r_H < \infty$, the same full-radius step gives $F(x) \ge r_H g$ and therefore the stronger bound $g^2 \le F(x)^2/r_H^2$. If $H_x = 0$ and $r_H = \infty$, the upper model is $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle$ for every y. Choosing $y = x - t\nabla f(x)$ and letting $t \to \infty$ would force $f^\star = -\infty$ unless $g = 0$. Since a minimizer exists, $g = 0$. Thus all degenerate cases are covered. □

I.2 Runtime algorithm

Set as before $A_0 = D^2/\Delta$, $A_{\mathrm{tar}} = 6D^2/\Delta$, and for a user-chosen horizon T,

$$a = \frac{5D^2}{\Delta T}, \qquad A_t = A_0 + ta, \qquad \ell_\rho = \log\frac{8}{\rho}, \qquad \Lambda_{\Delta,T} = \frac{\Delta T}{128 D \ell_\rho}.$$

The runtime update is exactly Procedure 1, but Λ is replaced by $\Lambda_{\Delta,T}$ and no hidden model parameters are read.

Lemma 21 (Horizon-induced localization)**.** *If $\Phi_{t-1} \le 3D^2$ and*

$$T \ge 40D\sqrt{\frac{\Gamma_\Delta}{\Delta}}, \tag{65}$$

then $F(y_{t-1}) \le 3\Delta$, $F(u_t) < 4\Delta$, both local-model displacements are at most r_H, $\mathcal{H}(u_t)a^2/A_t \le 1/64$, and

$$\|\nabla f(u_t)\| \le 4\sqrt{\Delta\Gamma_\Delta}.$$

Proof. Fix $t \in \{1, \dots, T\}$ and assume $\Phi_{t-1} \le 3D^2$ together with the horizon condition equation 65. We first control the previous output gap and both accelerated displacements, then apply the local upper model to the query point, and finally verify the quadratic coefficient and the query-gradient bound.

Output gap and displacements. The potential bound and $A_{t-1} \ge A_0 = D^2/\Delta$ first give

$$F(y_{t-1}) \le \frac{\Phi_{t-1}}{A_{t-1}} \le \frac{3D^2}{D^2/\Delta} = 3\Delta.$$

Next, $a/A_t \le a/A_0 = 5/T$. Since projection keeps every z_j in Q, and each y_j is a convex combination of points in Q, all y_j, z_j belong to Q. Therefore

$$\|u_t - y_{t-1}\| = \frac{a}{A_t}\|z_{t-1} - y_{t-1}\| \le \frac{5D}{T}, \tag{66}$$

$$\|y_t - u_t\| = \frac{a}{A_t}\|z_t - z_{t-1}\| \le \frac{5D}{T}. \tag{67}$$

Under equation 65,

$$\frac{5D}{T} \le \frac{1}{8}\sqrt{\frac{\Delta}{\Gamma_\Delta}}.$$

Because $\Gamma_\Delta \ge 4\Delta/r_H^2$, the last quantity is at most $r_H/16$. Thus both displacements in equation 66–equation 67 are within the radius of Assumption 2.

Query gap. At y_{t-1}, Lemma 20 and $F(y_{t-1}) \le 3\Delta$ give

$$\|\nabla f(y_{t-1})\|^2 \le 4F(y_{t-1})\left(H_0 + H_1 F(y_{t-1}) + \frac{F(y_{t-1})}{r_H^2}\right) \le 12\Delta\Gamma_\Delta. \tag{68}$$

Apply the local upper model from y_{t-1} to u_t. Using equation 66, equation 68, and $\mathcal{H}(y_{t-1}) \le \Gamma_\Delta$, we obtain

$$\begin{aligned} F(u_t) &\le F(y_{t-1}) + \|\nabla f(y_{t-1})\|\|u_t - y_{t-1}\| + \mathcal{H}(y_{t-1})\|u_t - y_{t-1}\|^2 \\ &\le 3\Delta + \sqrt{12\Delta\Gamma_\Delta}\frac{1}{8}\sqrt{\frac{\Delta}{\Gamma_\Delta}} + \Gamma_\Delta\frac{1}{64}\frac{\Delta}{\Gamma_\Delta} = \left(3 + \frac{\sqrt{12}}{8} + \frac{1}{64}\right)\Delta < 4\Delta. \end{aligned}$$

Consequently, $\mathcal{H}(u_t) = H_0 + H_1 F(u_t) \le H_0 + 4H_1\Delta \le \Gamma_\Delta$.

Quadratic coefficient and query gradient. Finally,

$$\frac{a^2}{A_t} \le \frac{a^2}{A_0} = \frac{25D^2}{\Delta T^2}.$$

Combining this inequality with $T^2 \ge 1600D^2\Gamma_\Delta/\Delta$ gives

$$\frac{\mathcal{H}(u_t)a^2}{A_t} \le \frac{25D^2\Gamma_\Delta}{\Delta T^2} \le \frac{1}{64}.$$

Applying Lemma 20 once more, now at u_t, and using $F(u_t) < 4\Delta$, yields

$$\|\nabla f(u_t)\|^2 \le 4F(u_t)\Gamma_\Delta < 16\Delta\Gamma_\Delta.$$

Taking square roots proves the stated gradient bound and completes the localization argument. □

Assume the actual but unknown oracle parameters obey Assumption 3. Define only for analysis

$$\bar{\varsigma}_{p,\Delta}^p := \sigma_0^p + \left(4\sigma_1\sqrt{\Delta\Gamma_\Delta}\right)^p + \sigma_2^p(4\Delta)^{p/2}. \tag{69}$$

Theorem 22 (Horizon-driven parameter-free contraction)**.** *The runtime trajectory described above does not use p, σ_i, or H_i. Nevertheless, if the actual parameters satisfy the assumptions and*

$$T \ge \max\left\{1024\ell_\rho D\sqrt{\frac{\Gamma_\Delta}{\Delta}},\ 128\,160^{1/(p-1)}\ell_\rho\left(\frac{\bar{\varsigma}_{p,\Delta}D}{\Delta}\right)^q\right\}, \tag{70}$$

then $\mathbb{P}\{F(y_T) \le \Delta/2\} \ge 1 - \rho$.

Proof. Fix the phase data $(w, \Delta, Q, D, T, \rho)$ and the filtration from Assumption 3. Recall that $a = 5D^2/(\Delta T)$, $A_t = D^2/\Delta + ta$, and $\Lambda_{\Delta,T} = \Delta T/(128D\ell_\rho)$. At iteration t, write

$$\beta_t := \mathbb{E}_t\widehat{g}_t - \nabla f(u_t), \qquad \zeta_t := \widehat{g}_t - \mathbb{E}_t\widehat{g}_t, \qquad v_t := \mathbb{E}_t\|\zeta_t\|^2.$$

The proof first stops the potential before localization can fail, then verifies the clipping bias and conditional moment bounds on the stopped path, applies Freedman's inequality to the two centered channels, and finally removes the stopping time by contradiction.

Stopped process and localization. Set

$$\tau := \inf\{t \in \{1, \dots, T\} : \Phi_t > 3D^2\}, \qquad I_t := \mathbf{1}\{t - 1 < \tau\},$$

with $\inf \varnothing = T + 1$. The potential Φ_j is $\mathcal{F}_j$-measurable, and

$$\{t - 1 < \tau\} = \bigcap_{j=1}^{t-1} \{\Phi_j \le 3D^2\} \in \mathcal{F}_{t-1}.$$

Therefore I_t is predictable. Moreover, $I_t = 1$ implies that the localization premise $\Phi_{t-1} \le 3D^2$ holds. The first condition in equation 70 is stronger than equation 65, so Lemma 21 applies before τ.

Clipping radius and conditional moment bound. The same horizon condition also makes the clipping radius large enough. Indeed,

$$2\|\nabla f(u_t)\| \le 8\sqrt{\Delta\Gamma_\Delta} \le \frac{\Delta T}{128D\ell_\rho} = \Lambda_{\Delta,T},$$

where the second inequality is exactly $T \ge 1024\ell_\rho D\sqrt{\Gamma_\Delta/\Delta}$ after rearrangement. Hence $\Lambda_{\Delta,T} \ge 2\|\nabla f(u_t)\|$, and the proof of Lemma 6 applies with the localized moment scale $\bar{\varsigma}_{p,\Delta}$ from equation 69.

Clipping bias and centered second moment. The second horizon condition implies

$$\Lambda_{\Delta,T}^{p-1} \ge 160\bar{\varsigma}_{p,\Delta}^p \frac{D}{\Delta}. \tag{71}$$

To verify equation 71, the second term in equation 70 and the definition of the clipping radius give

$$\Lambda_{\Delta,T} = \frac{\Delta T}{128D\ell_\rho} \ge 160^{1/(p-1)} \frac{\Delta}{D} \left(\frac{\bar{\varsigma}_{p,\Delta} D}{\Delta}\right)^q = \left(160\bar{\varsigma}_{p,\Delta}^p \frac{D}{\Delta}\right)^{1/(p-1)}.$$

The last equality uses $q = p/(p-1)$, so raising both sides to the positive power $p - 1$ proves equation 71. Lemma 6 and equation 71 now give, before τ,

$$\|\beta_t\| \le 2\bar{\varsigma}_{p,\Delta}^p \Lambda_{\Delta,T}^{1-p} \le \frac{\Delta}{80D},$$

and

$$v_t := \mathbb{E}_t\|\zeta_t\|^2 \le 2\bar{\varsigma}_{p,\Delta}^p \Lambda_{\Delta,T}^{2-p} \le \frac{\Lambda_{\Delta,T}\Delta}{80D}.$$

Stopped potential recursion. Lemma 21 gives $\mathcal{H}(u_t)a^2/A_t \le 1/64 < 1/16$, so the coefficient condition of Lemma 7 holds. The two displacements are within r_H by Lemma 21, so both uses of the local upper model in that lemma are also valid. Multiplying its pathwise recursion by the predictable indicator I_t, using $\|\zeta_t + \beta_t\|^2 \le 2\|\zeta_t\|^2 + 2\|\beta_t\|^2$, and adding and subtracting v_t gives exactly the stopped recursion equation 30.

Predictable contributions. We now verify each numerical channel for the new threshold. Since $Ta = 5D^2/\Delta$,

$$\sum_{t=1}^{T} aI_t D\|\beta_t\| \le TaD\frac{\Delta}{80D} = \frac{D^2}{16},$$

while

$$2\sum_{t=1}^{T} a^2 I_t\|\beta_t\|^2 \le 2Ta^2\frac{\Delta^2}{80^2D^2} = \frac{D^2}{128T}.$$

For the conditional-variance contribution, use $\Lambda_{\Delta,T} = \Delta T/(128D\ell_\rho)$:

$$2\sum_{t=1}^{T} a^2 I_t v_t \le 2Ta^2 \frac{\Lambda_{\Delta,T}\Delta}{80D} = \frac{5D^2}{1024\ell_\rho} < \frac{D^2}{400},$$

where the last inequality follows from $\ell_\rho = \log(8/\rho) > \log 8$.

Martingale terms. For $j \le T$, define

$$M_j^{(1)} := -\sum_{t=1}^{j} aI_t \langle \zeta_t, z_{t-1} - x^\star \rangle, \qquad M_j^{(2)} := 2a^2 \sum_{t=1}^{j} I_t(\|\zeta_t\|^2 - v_t).$$

The point u_t, the iterate z_{t-1}, and the indicator I_t are $\mathcal{F}_{t-1}$-measurable. By construction, $\mathbb{E}_t\zeta_t = 0$ and $\mathbb{E}_t(\|\zeta_t\|^2 - v_t) = 0$. Thus both displayed processes are scalar martingales adapted to $(\mathcal{F}_j)_{j\ge 0}$. Moreover, clipping gives $\|\widehat{g}_t\| \le \Lambda_{\Delta,T}$ almost surely and Jensen's inequality gives $\|\mathbb{E}_t\widehat{g}_t\| \le \Lambda_{\Delta,T}$; hence $\|\zeta_t\| \le 2\Lambda_{\Delta,T}$ almost surely.

For the signed martingale, $\|z_{t-1} - x^\star\| \le D$ and the localized bound on v_t imply

$$|\Delta M_t^{(1)}| \le 2a\Lambda_{\Delta,T}D = \frac{5D^2}{64\ell_\rho},$$

$$V_T^{(1)} := \sum_{t=1}^{T} \mathbb{E}_t[(\Delta M_t^{(1)})^2] \le a^2D^2 \sum_{t=1}^{T} I_t v_t \le a^2D^2T\frac{\Lambda_{\Delta,T}\Delta}{80D} = \frac{5D^4}{2048\ell_\rho}.$$

For the quadratic martingale, $|\|\zeta_t\|^2 - v_t| \le 4\Lambda_{\Delta,T}^2$ and $\|\zeta_t\|^4 \le 4\Lambda_{\Delta,T}^2\|\zeta_t\|^2$ almost surely. Consequently,

$$|\Delta M_t^{(2)}| \le 8a^2\Lambda_{\Delta,T}^2 = \frac{25D^2}{2048\ell_\rho^2},$$

$$V_T^{(2)} := \sum_{t=1}^{T} \mathbb{E}_t[(\Delta M_t^{(2)})^2] \le 16a^4\Lambda_{\Delta,T}^2 \sum_{t=1}^{T} I_t v_t \le 16a^4\Lambda_{\Delta,T}^2 T\frac{\Lambda_{\Delta,T}\Delta}{80D} = \frac{125D^4}{2097152\ell_\rho^3}.$$

Freedman bound and one-phase contraction. We apply the two-sided maximal Freedman inequality stated in Appendix B with $u = D^2/8$. For the signed martingale, the denominator in the Freedman exponent satisfies

$$2\left(V_T^{(1)} + \frac{L_1 u}{3}\right) \le \frac{2D^4}{\ell_\rho}\left(\frac{5}{2048} + \frac{5}{1536}\right) = \frac{35D^4}{3072\ell_\rho} \le \frac{D^4}{64\ell_\rho} = \frac{u^2}{\ell_\rho}.$$

For the quadratic martingale, $\ell_\rho > 1$ gives

$$2\left(V_T^{(2)} + \frac{L_2 u}{3}\right) \le \frac{2D^4}{\ell_\rho^2}\left(\frac{125}{2097152} + \frac{25}{49152}\right) \le \frac{D^4}{64\ell_\rho} = \frac{u^2}{\ell_\rho}.$$

The second numerical inequality holds because $2(125/2097152 + 25/49152) < 1/64$. Thus the exponent $u^2/[2(V + Lu/3)]$ is at least ℓ_ρ for each martingale, and each failure probability is at most $2e^{-\ell_\rho} = \rho/4$. On their common good event, which has probability at least $1 - \rho/2$, the stopped recursion and equation 21 give, uniformly in $j \le T$,

$$\Phi_{j\wedge\tau} \le \frac{3}{2}D^2 + \frac{D^2}{16} + \frac{D^2}{128T} + \frac{D^2}{400} + \frac{D^2}{8} + \frac{D^2}{8} < 3D^2,$$

where the terms after $3D^2/2$ are, respectively, the accumulated bias inner product, squared bias, predictable variance, and the two martingale deviations. If $\tau \le T$, choosing $j = \tau$ contradicts the definition of τ. Hence $\tau = T + 1$ on the good event. Finally, $A_T = A_0 + Ta = 6D^2/\Delta$ and $A_T F(y_T) \le \Phi_T \le 3D^2$, so $F(y_T) \le \Delta/2$. Since the good event has probability at least $1 - \rho/2 \ge 1 - \rho$, the theorem follows. □

Remark 23 (Meaning of "parameter-free"). *Theorem 22 is fixed-budget adaptation. The unknown parameters determine which horizons satisfy equation 70; the algorithm does not certify from the observed trajectory that a particular unknown sufficient horizon has already been reached. A common-budget restart wrapper can allocate $T = \lfloor N/S \rfloor$ to every phase, losing at most another factor S relative to phase-wise tuning. In the strongly convex wrapper, μ remains an operational parameter because it determines the shrinking sets Q_s.*

J Experimental Details and Additional Results

Construction. We sample 64 vehicle and 64 animal images from the CIFAR-10 training set and extract frozen DINOv2 ViT-S/14 features. The encoder is never updated; optimization acts only on a convex kernel head, the dual counterpart of last-layer fitting on fixed representations. After row normalization, let G be the feature Gram matrix. We choose a ridge $\gamma \geq 0$ and a scalar $c > 0$ so that $K = (G + \gamma I)/c$ has largest eigenvalue one and condition number 20. The labels are $y_i = 1$ for vehicles and $y_i = -1$ for animals. For fixed weights $\omega_i > 0$, the component objective and gradient are

$$f_i(\alpha) = \omega_i[\cosh((K\alpha)_i - y_i) - 1], \qquad g_i(\alpha) = \omega_i \sinh((K\alpha)_i - y_i)K_{i:}.$$

Because K is invertible, $\alpha^\star = K^{-1}y$ gives $f_i(\alpha^\star) = 0$ and $g_i(\alpha^\star) = 0$ for every i. Furthermore,

$$\nabla^2 f(\alpha) = \frac{1}{n} K^\top \operatorname{diag}\big(\omega_i \cosh((K\alpha)_i - y_i)\big)K,$$

so f is $\mu_{\exp}$-strongly convex with $\mu_{\exp} := \min_i \omega_i \lambda_{\min}(K)^2/n$, and $\|\nabla^2 f(\alpha)\| \leq R^2(1 + f(\alpha))$ for $R := \max_i \|K_{i:}\|$ because $n^{-1}\sum_i \omega_i = 1$. Thus the experiment has exact interpolation and an (H_0, H_1) Hessian bound with $H_0 = H_1 = R^2$.

In the equal-weight regime, $\omega_i = 1$. In the power-law regime, the weights are drawn once from a Pareto distribution with shape 1.5, normalized to mean one, and then stored as fixed data. Their minimum, median, 99th percentile, and maximum are 0.392, 0.590, 7.88, and 9.75, respectively. At initialization, the descriptive Hill estimate for the tail index of component-gradient norms is 1.52. This is an empirical tail diagnostic only: a finite dataset does not generate a literally infinite-variance oracle.

Protocol and tuning. All methods start from zero and use the same streams of uniformly sampled component indices. Since every update uses one component gradient, iteration count and oracle calls coincide. Each method is tuned on seeds 101–106 and evaluated on the disjoint seeds 20001–20030. We choose among 16 configurations by final 90th-percentile relative gap, then by the median, and fix the selected configuration before evaluation. The budget is $8{,}192$, the desired absolute gap is 10^{-5}, and RCSAG uses 16 equal-length restart phases. Its step multiplier ranges over $\{0.03, 0.1, 0.3, 1\}$ and its fixed-budget clipping multiplier over $\{64, 512, 4096, 32768\}$. Clipped SGD follows Algorithm 2 of Gorbunov et al. (2020); its step parameter ranges over $\{0.1, 0.3, 1, 1.8\}/L_{\text{init}}$ and its clipping radius over $\{0.25, 1, 4, 16\}$ times the median initial component-gradient norm. For both SSTM variants, the weight parameter ranges over $\{0.5, 2, 8, 32\}N$ and the clipping parameter over $\{0.025, 0.1, 0.4, 1.6\}$. These are practical tuning grids; they do not attempt to instantiate the conservative numerical constants of Theorem 8.

Table 2: Held-out results after $8{,}192$ component-gradient calls. N_{90} is the first checkpoint after which at least 90% of runs remain below 10^{-5}; "–" means that no such checkpoint exists.

	Equal weights		Power-law weights		
Method	Median gap	N_{90}	Median gap	N_{90}	Clipped
RCSAG	3.47×10^{-7}	6,705	1.22×10^{-6}	8,192	0.15%
Clipped SGD	1.58×10^{-4}	–	2.95×10^{-4}	–	7.42%
Clipped-SSTM	4.14×10^{-4}	–	6.10×10^{-4}	–	2.23%
R-clipped-SSTM	3.21×10^{-2}	–	3.32×10^{-2}	–	60.51%

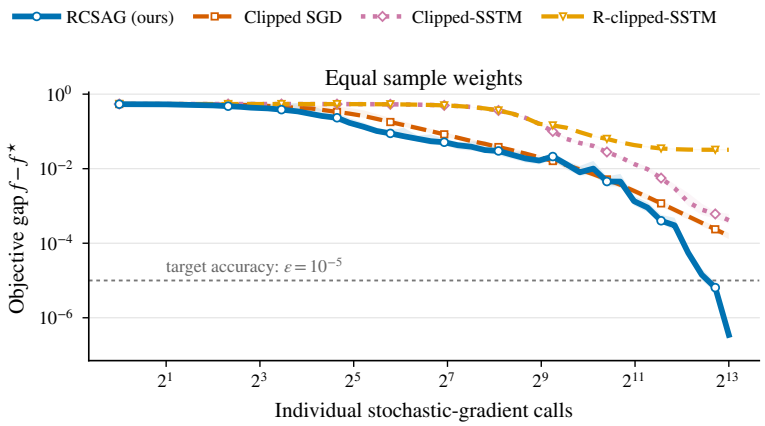


Figure 3: Equal-weight control: median gap with 10th–90th percentile bands over 30 held-out seeds.

Additional observations. The equal-weight control in Figure 3 preserves the same ranking: RCSAG reaches 10^{-5} with sustained 90% success after 6,705 calls. Under fixed Pareto weights, its median gap is 1.22×10^{-6} versus 2.95×10^{-4} for Clipped SGD. The median clipped fraction of RCSAG changes from zero to only 0.15%, so the method clips rare large components rather than most updates.

Synthetic finite-p check. We also rerun the comparison on the eight-dimensional strongly convex objective $f(x) = \sum_{j=1}^{8} \lambda_j b^{-2}[\cosh(bx_j) - 1]$, where $b = 1.5$ and λ_j are geometrically spaced in $[1, 16]$. The oracle adds radial Student-t noise with 1.8 degrees of freedom and scale $(\sigma_0^p + \sigma_2^p F(x)^{p/2})^{1/p}$, using $p = 1.5$, $\sigma_0 = 0.1$, and $\sigma_2 = 0.4904$. Each method is selected from 16 configurations on six tuning seeds and evaluated on 300 disjoint seeds with 16,384 individual gradient calls.

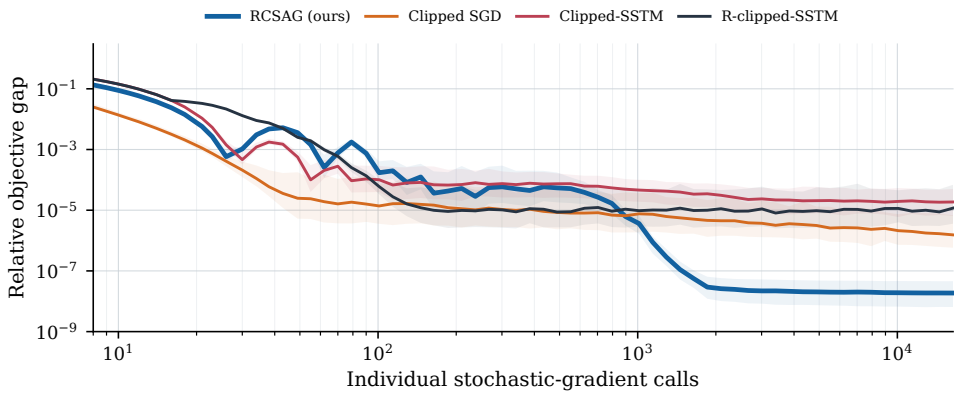


Figure 4: Synthetic Student-t experiment: median relative gap with 10th–90th percentile bands over 300 held-out seeds.

RCSAG reaches relative gap 10^{-6} with sustained 90% success after 1,290 calls and succeeds on all 300 runs by the final budget. Its final 90th-percentile gap is 4.60×10^{-8}, compared with 6.86×10^{-6} for Clipped SGD; neither SSTM variant reaches the target. The median fraction of clipped RCSAG updates is 10.1%. As above, these runs use tuned practical constants and do not certify the theorem's numerical constants.